Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Estudi de la demanda de la instal·lació fotovoltaica aïllada de l'ETSEIB

MEMÒRIA

Autora: Anaïs Ferrara Marzo
Director: Roberto Villafáfila Robles

Convocatòria: Juny 2019

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

L'objectiu d'aquest projecte és trobar el millor model per fer una predicció de la demanda elèctrica de la biblioteca de l'ETSEIB. Els 66 endolls de la sala d'estudi estan alimentats amb l'energia elèctrica procedent de la instal·lació fotovoltaica aïllada que conté 16 panells solars fotovoltaics a la coberta de l'escola.

En primer lloc, es fa un estudi de l'estat de l'art de l'energia solar fotovoltaica, explicant com s'obté, els tipus de instal·lacions existents i els components principals presents en una instal·lació solar fotovoltaica. També s'analitza la situació actual de l'anàlisi predictiu: què és, aplicacions que té, tipus de model i algunes de les principals tècniques que s'apliquen. Això permet al lector assentar les bases de coneixement per entendre plenament el projecte.

Seguidament es fa una descripció del funcionament de la instal·lació solar fotovoltaica de l'ETSEIB i dels principals components que la formen. També s'indiquen els dispositius d'anàlisi de dades que comprèn, així com la plataforma mitjançant la qual es poden obtenir aquestes dades.

Abans de passar a la part pràctica del treball és necessari explicar la metodologia general emprada a l'hora de fer una predicció, la qual s'aplica en aquest projecte. L'estructura principal consisteix en la descàrrega, anàlisi i tractament de les dades, la comparació de la qualitat que ens aporten diferents algoritmes, l'elecció d'un d'ells i l'optimització d'aquest.

Per tal de realitzar la predicció de la demanda elèctrica dels llocs de treball de l'ETSEIB es fa ús de Scikit Learn, la principal llibreria per estudiar aprenentatge computacional amb el llenguatge Python.

Els resultats dels diferents passos es mostren emprant taules i gràfics, cosa que fa que la comparació de models i la comprensió dels resultats finals sigui entenedora. També es redacten totes les decisions importants preses durant el transcurs del treball, per tal que el lector pugui seguir el fil fins arribar al model final.

Per últim, es presenten els models finals escollits i les limitacions que aquests presenten.

Sumari

1. Prefaci

1.1. Motivació

L'energia és el motor de les economies industrialitzades del planeta. Avui en dia és impossible imaginar un país avançat sense una àmplia i moderna xarxa de creació i distribució d'aquest recurs tan important. Pot provenir de diverses fonts, tant d'energies renovables (eòlica, solar, hidràulica...) com de no renovables (carbó, gas, petroli...) i les seves plantes poden estar ubicades a diferents zones geogràfiques. Per tant, és imprescindible una gran coordinació per garantir el proveïment tant a la demanda residencial, com a la comercial i industrial.

Tot i així, un dels majors inconvenients que la generació de l'energia elèctrica comporta és la incapacitat d'emmagatzemar grans quantitats d'aquesta d'una manera eficient. Degut a aquest fet fonamental, no són pocs els països que tracten constantment d'optimitzar els seus sistemes de previsió de la demanda d'energia per aconseguir fer-los més precisos i fiables. La millora d'aquests models de predicció no només seria beneficiós a nivell econòmic, sinó també suposaria una important millora a nivell mediambiental i social, ja que limitaria les emissions i ajudaria a satisfer més eficientment les necessitats de tots els consumidors. [16]

Per tots aquests motius, tots els grans implicats del mercat elèctric han treballat conjuntament per conformar models que busquin ajustar de la manera més eficient la demanda energètica dels consumidors amb l'oferta de producció disponible.

Per veure un exemple del seguiment de la demanda d'energia elèctrica a nivell estatal, es pot accedir al servidor web de la Red Eléctrica de España [28]. A l'apartat corresponent es pot visualitzar un gràfic de sèrie temporal en el que s'indica la demanda prevista, la demanda programada i la demanda real del dia que s'indiqui. A la Figura 1 es mostra l'exemple del dia 19/05/2019.

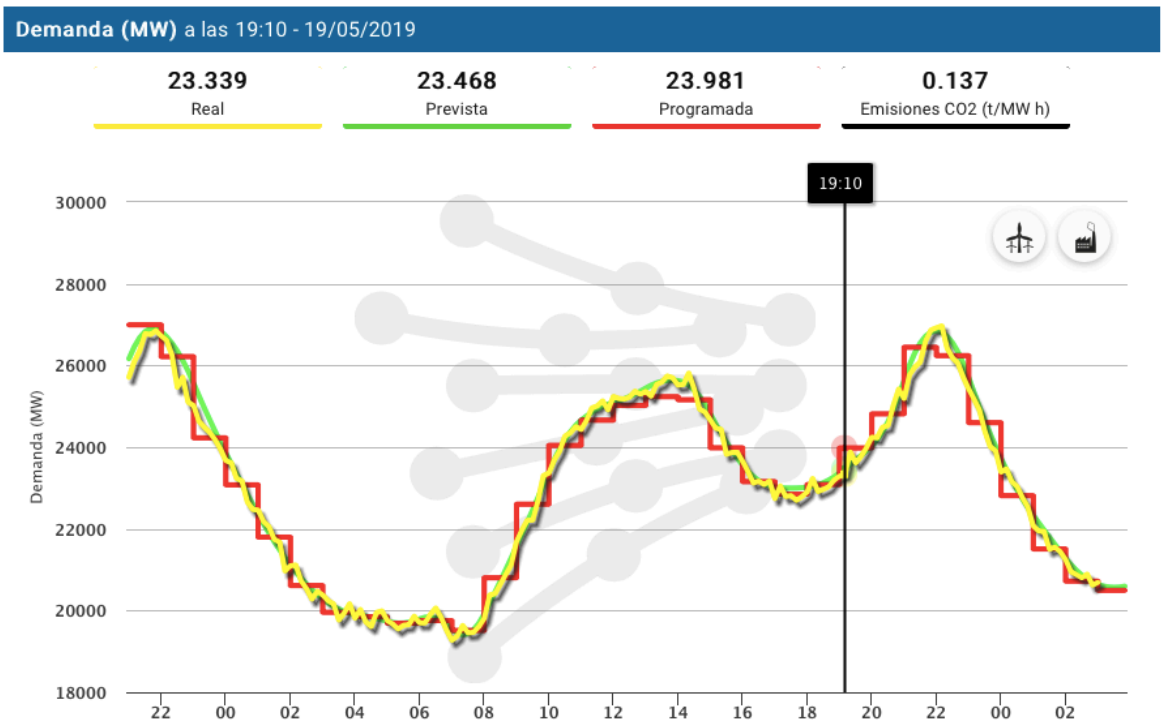

Figura 1. Seguiment de la demanda elèctrica a Espanya el dia 19/05/2019 [28]

De totes maneres, aquesta gestió de l'energia elèctrica també es pot fer a escala més reduïda, com per exemple el d'una instal·lació fotovoltaica aïllada. Aquest és el cas del present treball, que s'endinsa en el món de la previsió de la demanda elèctrica per tal de, en un futur, utilitzar els resultats per gestionar les bateries de la instal·lació fotovoltaica de forma més adequada.

Què vol dir exactament gestionar les bateries de forma més adequada? Doncs optimitzar la seva gestió per tal d'aprofitar al màxim l'energia que es podria generar amb la radiació solar. D'aquesta manera es podria aconseguir que les bateries mai es quedessin sense energia suficient per respondre a la demanda. Fins i tot es podria pensar en sobredimensionar les instal·lacions i alimentar encara més endolls o aparells.

Una altra aplicació d'aquesta predicció podria ser descobrir les èpoques o moments de la setmana en els que hi ha menys demanda i així decidir quan realitzar accions de manteniment en la instal·lació. Pel mateix motiu, es podria descobrir quin tant per cent dels panells fotovoltaics es pot desconnectar en un moment concret si es vol seguir responent a la demanda de la biblioteca.

1.2. Treballs anteriors

Aquest treball es presenta com la continuació a una sèrie de treballs de fi de grau que s'han realitzat a l'ETSEIB en els darrers anys. Tots ells relacionats amb la instal·lació fotovoltaica que alimenta 66 espais de treball de la biblioteca de l'escola, cadascun dels quals disposa d'un endoll amb els que es carreguen principalment mòbils, *tablets* i portàtils.

El primer projecte va permetre el muntatge a mitjans de 2017 de la instal·lació solar fotovoltaica a partir d'un conjunt de components de la marca Circutor i 16 panells solars situats a la coberta de la biblioteca de l'escola. [25]

El segon projecte, realitzat a mitjans del 2018, va realitzar una pàgina web on es mostra la informació a temps real de diferents paràmetres del sistema. Això implica que es van habilitar diferents punts de recol·lecció d'informació dels diferents elements connectats. [22]

El projecte més recent es va presentar el gener del 2019 i feia una avaluació del funcionament i de l'estat actual del sistema fotovoltaic aïllat partint de les dades emmagatzemades gràcies al projecte anterior. [29]

El present treball de fi de grau pretén continuar amb l'estudi de les dades disponibles i entrar en el món de la previsió de la demanda elèctrica. Concretament, el que pretén el present projecte és trobar el model que descrigui millor la demanda elèctrica de la instal·lació solar fotovoltaica de l'ETSEIB per tal de poder-ne fer una predicció.

2. Introducció

El present treball realitza un estudi de la demanda elèctrica de la instal·lació fotovoltaica que alimenta 66 espais de treball de la biblioteca de l'ETSEIB, cadascun dels quals disposa d'un endoll amb els que es carreguen principalment mòbils, *tablets* i portàtils.

Cal tenir present que es tracta d'un comportament de la demanda elèctrica molt particular, ja que està íntimament relacionada amb el calendari acadèmic dels estudiants de l'escola. Per tant, les tendències estudiades per les demandes elèctriques domèstiques o industrials no són d'aplicació en aquest cas.

La base de coneixement del present projecte és l'aprenentatge computacional, un camp de la intel·ligència artificial que es dedica al disseny, anàlisi i desenvolupament d'algoritmes que permeten que les màquines evolucionin. A partir de la implementació d'aquests, es desitja trobar el model que faci una predicció més precisa i exacta de la demanda.

L'estudi es farà relacionant les dades rellevants de la instal·lació fotovoltaica aïllada amb el calendari acadèmic de l'ETSEIB, l'època de l'any i altres variables que es considerin d'influència.

2.1. Objectius

L'objectiu principal és fer una predicció de la demanda elèctrica de la instal·lació fotovoltaica aïllada de l'ETSEIB fent ús d'algoritmes d'aprenentatge computacional.

Les tasques a realitzar per assolir aquest objectiu es redacten a continuació:

- Situar-se en l'àmbit de l'energia fotovoltaica. Donat que es tracta d'una instal·lació aïllada i hi ha una dependència molt gran amb el comportament d'aquesta, és interessant entendre com s'aconsegueix l'energia elèctrica que satisfà la demanda estudiada.
- Estudiar l'estat de l'art de l'anàlisi predictiu: entendre què és, quan pot ser d'utilitat i quins són els models i tècniques aplicables.
- Fer una descripció general de la instal·lació fotovoltaica de l'ETSEIB, en la que quedin especificats els diferents dispositius de la instal·lació fotovoltaica i de la monitorització de dades.
- Veure la metodologia general seguida per fer una predicció.
- Identificar les característiques principals de la predicció de la demanda elèctrica de les taules d'estudi de la biblioteca de l'ETSEIB.
- Programar i arribar a resultats per després analitzar-los i entendre'ls.

2.2. Abast

El present treball realitza una predicció de la demanda elèctrica de la instal·lació fotovoltaica que alimenta 66 espais de treball de la biblioteca de l'ETSEIB, cadascun amb un sol endoll.

Aquesta predicció es realitza fent ús d'algoritmes d'aprenentatge computacional, concretament, d'aquells disponibles per la llibreria de Scikit Learn de Python.

Les dades de les que es parteix per fer la predicció són aquelles obtingudes amb els dispositius de monitorització de dades que disposa la instal·lació fotovoltaica de l'ETSEIB. També es fa ús del calendari acadèmic de l'escola i el calendari de la biblioteca. Concretament, a causa de la limitació d'emmagatzematge de dades dels dispositius, s'utilitzen dades que van des del 20/03/18 al 21/05/19, període que inclou poc més de dos quadrimestres de l'escola.

3. Estat de l'art de l'energia fotovoltaica

3.1. L'energia fotovoltaica

S'entén per energia fotovoltaica com la transformació directa de la radiació solar en electricitat. Aquesta transformació es produeix en uns dispositius anomenats panells fotovoltaics. En ells la radiació solar excita els electrons d'un dispositiu semiconductor generant una petita diferència de potencial. La connexió en sèrie d'aquests dispositius permet obtenir diferències de potencial més grans.

Encara que l'efecte fotovoltaic era conegut des del segle XIX, va ser a la dècada dels 50, en plena carrera espacial, quan els panells fotovoltaics van començar a experimentar un important desenvolupament. Inicialment utilitzats per subministrar electricitat a satèl·lits geoestacionaris de comunicacions, avui en dia constitueixen una tecnologia de generació elèctrica renovable. [3]

Una de les principals virtuts de la tecnologia fotovoltaica és el seu aspecte modular, podent així construir des de gegants plantes fotovoltaiques al sòl fins a petits panells per teulades.

3.2. Generació d'energia solar fotovoltaica

La generació de l'energia fotovoltaica es fa en les anomenades cèl·lules solars. Aquestes basen el seu funcionament en l'efecte fotovoltaic, convertint directament en electricitat els fotons provinents de la llum del Sol.

Una cèl·lula solar es comporta com un díode, els quals estan formats per una capa de semiconductor de tipus n i una altra de tipus p. La capa exposada a la radiació solar és la de tipus n i la situada a la zona fosca és la de tipus p. Això produeix una diferència de voltatge o de potencial entre les dues capes que és capaç de conduir una corrent a través d'un circuit extern, de manera que es pot produir treball útil.

El material més utilitzat per fabricar aquestes cèl·lules solars és el silici, tot i que també s'implementen d'altres, com el germani.

3.3. Classificació de les instal·lacions solars fotovoltaiques

La classificació de les instal·lacions solars fotovoltaiques (ISF) es pot realitzar en funció de l'aplicació a la que estan destinades. D'aquesta manera, es diferencia entre aplicacions autònomes i aplicacions connectades a la xarxa. [1]

3.3.1. Aplicacions autònomes

Produeixen electricitat sense cap tipus de connexió a la xarxa elèctrica per tal de donar energia al lloc on es troben ubicades. Es poden diferenciar entre les aplicacions espacials i les terrestres. De les aplicacions terrestres cal destacar-ne algunes com telecomunicacions, electrificació de zones rurals i aïllades, senyalització, l'enllumenat públic, bombeig d'aigua, xarxes VSAT, telemesura, …

3.3.2. Aplicacions connectades a la xarxa

En aquests casos el productor no utilitza l'energia directament, sinó que es ven a l'organisme encarregat de la gestió de l'energia en el país. Cal distingir entre:

- Centrals fotovoltaiques i horts solars:
  Es tracta de recintes en els que es concentren un número determinat d'instal·lacions fotovoltaiques de diferents propietaris amb la finalitat de vendre l'electricitat produïda a la companyia elèctrica amb la qual s'hagi establert contracte.
- Edificis fotovoltaics:
  Es tracta d'una de les últimes aplicacions desenvolupades per l'ús de l'energia fotovoltaica i consisteix en combinar la doble funció, com element constructiu i com a productor d'electricitat, dels mòduls fotovoltaics.

3.4. Elements d'una ISF

De manera general, una ISF s'ajusta a un esquema com el mostrat a la Figura 2. Seguidament es procedirà a explicar el funcionament de cadascun d'aquests elements. [13]

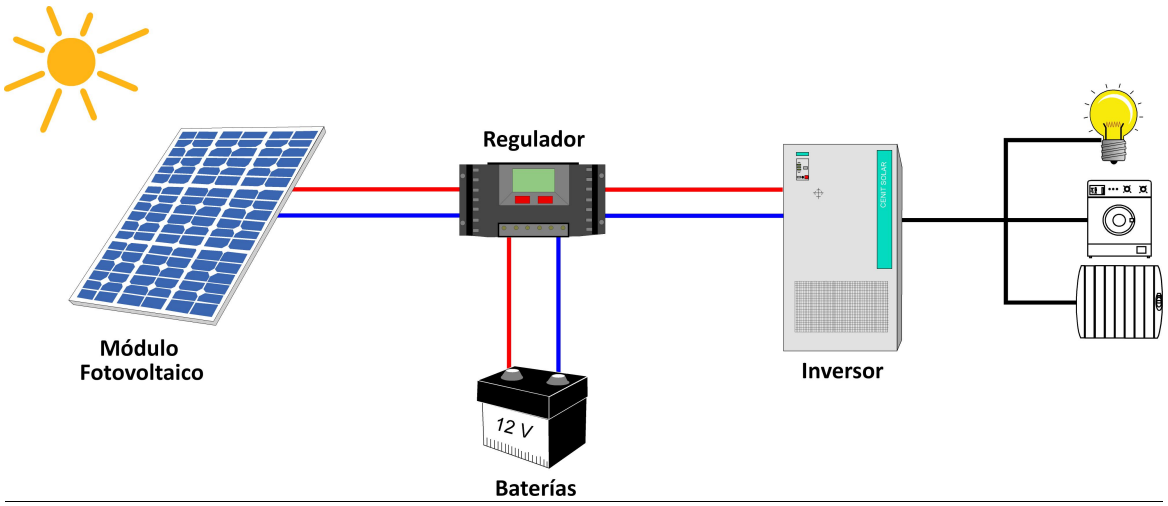

Figura 2. Esquema bàsic d'una instal·lació solar fotovoltaica [6]

3.4.1. Mòdul fotovoltaic o panell solar.

Abans de poder definir què és el mòdul fotovoltaic cal parlar de les cèl·lules solars, que són l'element principal de qualsevol instal·lació d'energia solar. Fan de generador i converteixen directament en electricitat els fotons provinents de la llum del Sol. El seu funcionament es basa en l'efecte fotovoltaic.

El panell solar està format per un conjunt de cèl·lules connectades elèctricament, encapsulades i muntades sobre una estructura de suport o marc. Proporciona a la seva sortida de connexió una tensió contínua i es dissenya per uns valors concrets de tensió en els que treballarà el sistema fotovoltaic.

Hi ha la possibilitat d'utilitzar un sol panell o, en cas de necessitar una potència elevada, un conjunt de panells que es muntaran agrupats i connectats entre sí elèctricament.

Els tipus de panells solars venen donats per la tecnologia de fabricació de les cèl·lules i són fonamentalment:

- Silici cristal·lí (monocristal·lí o policristal·lí)
- Silici amorf

A la Taula 1 es poden veure els seus rendiments i algunes característiques principals:

Cèl·lules	**Silici**	**Rendiment laboratori**	**Rendiment directe**	**Característiques**
	Monocristal·lí	24%	15 - 18%	Opció més popular del mercat. Garanteix nivells d'eficiència dignes en totes les condicions meteorològiques. També són les més cares. Generalment d'un color blau uniforme.
	Policristal·lí	19 - 20%	12 - 14%	Baix rendiment en condicions d'il·luminació baixa. Bloc de silici que té múltiples cristalls. Tenen forma quadrada i una superfície semblant a un mosaic.
	Amorf	16%	< 10%	Funciona amb llum difusa baixa, inclús en dies nuvolats. El seu rendiment es redueix amb el temps. Làmina tallada a mida en la que s'observen tires primes que separen les cèl·lules.

Taula 1. Característiques principals dels tipus de panells solars [35]

3.4.2. Regulador de càrrega

La seva missió és evitar situacions de càrrega i descàrrega de la bateria amb la finalitat d'allargar la seva vida útil. És a dir, treballa en les dues zones. Per una banda assegura la càrrega suficient a l'acumulador i evita les situacions de sobrecàrrega i, per altra banda, s'ocupa d'assegurar el subministrament elèctric diari suficient i evitar la descàrrega excessiva de la bateria.

Cal mencionar que pel correcte funcionament de la instal·lació fotovoltaica, el mòduls solars tenen una tensió nominal major que la de les bateries. D'aquesta manera, si no existís el regulador, es podrien produir sobrecàrregues.

3.4.3. Acumulador: Bateria

L'arribada de l'energia solar als mòduls fotovoltaics no es produeix de manera uniforme, sinó que presenta variacions per diferents motius: duració de la nit, estacions de l'any, nuvolositat, etc. Aquest fet fa que sigui necessari utilitzar algun sistema que emmagatzemi l'energia per aquells moments en què la radiació rebuda sobre el generador fotovoltaic no sigui capaç de fer que la instal·lació funcioni en els valors dissenyats.

Per això serveixen les bateries, que són capaces de transformar l'energia química en elèctrica.

El funcionament d'una bateria en una instal·lació fotovoltaica és el mostrat a la Figura 3.

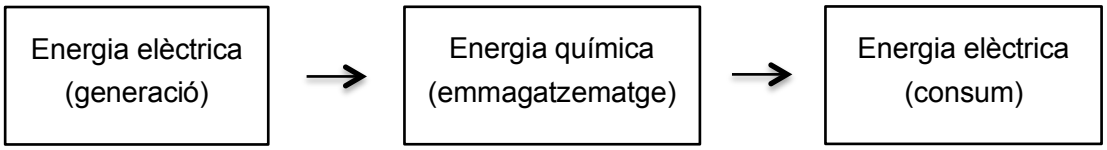

Figura 3. Esquema bàsic del funcionament de les bateries

Podem dir que les missions de les bateries en les instal·lacions fotovoltaiques són tres: emmagatzemar energia durant un determinat número de dies, proporcionar una potència instantània elevada i fixar la tensió de treball de la instal·lació.

3.4.4. Inversor

S'encarrega de convertir el corrent continu de la instal·lació en corrent altern, igual a l'utilitzat a la xarxa elèctrica: 220 V de valor eficaç i una freqüència de 50 Hz.

Les característiques desitjables per un inversor DC-AC es poden resumir de la següent manera: alta eficiència, baix consum al buit, alta fiabilitat, protecció contra curtcircuits, seguretat i bona regulació de la tensió i freqüència de sortida.

Cal comentar que en alguns casos els inversors funcionen també com a reguladors de càrrega de les bateries. En aquest cas no seria necessari incloure reguladors a la instal·lació.

4. Estat de l'art de l'anàlisi predictiu

En els últims anys s'ha posat especial atenció a l'anàlisi predictiu degut als avenços en la tecnologia que el recolza, especialment en les àrees de les dades massives (*Big Data)* i l'aprenentatge automàtic (*Machine Learning*).

Per exemple, una empresa com Aqualia, que dona un servei a una població de 27 milions de persones en una àrea tan fonamental com el subministrament d'aigua, utilitza l'anàlisi predictiu per preveure talls en el subministrament. Aqualia és capaç d'identificar patrons i tendències de consum amb la finalitat de predir la demanda en un escenari canviant tant per la climatologia com pels bruscs canvis de la població deguts al turisme. [2]

No només és útil en aquest camp, també s'aplica en altres àmbits, com el de la política. Es troba un exemple durant la campanya electoral que va enfrontar Barack Obama i Mitt Romney, en la que els sondejos mostraven un escenari en el que tots dos candidats alternaven el lideratge degut als *swing voters* (electors susceptibles de canviar el sentit del seu vot). La campanya d'Obama va decidir centrar-se en aquests votants identificant els grups que reaccionaven millor davant una trucada telefònica, l'enviament d'informació per correu o davant una visita a casa. Van arribar a registrar grups votants que podrien canviar el vot de forma negativa si se'ls contraataca. Ho van aconseguir gràcies a l'anàlisi predictiu. Es van aplicar tècniques de mineria de dades que van permetre identificar els patrons de comportament i crear un model. [21]

Empreses, administracions públiques i organitzacions troben en l'anàlisi predictiu una eina imprescindible que permet substituir les intuïcions i apreciacions personals per projeccions científiques capaces d'eliminar o disminuir les incerteses a l'hora d'elaborar les seves estratègies.

L'anàlisi predictiu és una subdisciplina de l'anàlisi de dades que utilitza tècniques d'estadística, com l'aprenentatge computacional o la mineria de dades, per desenvolupar models que preveuen esdeveniments futurs o conductes. Aquests models predictius permeten aprofitar els patrons de comportament trobats en les dades actuals i històriques per identificar riscos i oportunitats. [18]

4.1. Què és l'anàlisi predictiu

L'anàlisi predictiu és una àrea de la mineria de dades que consisteix en l'extracció d'informació existent en les dades i la seva utilització per predir tendències i patrons de comportament, podent aplicar-se sobre qualsevol esdeveniment desconegut, ja sigui en el passat, present o futur. L'anàlisi predictiu es fonamenta en la identificació de relacions entre

variables i esdeveniments passats, per després explotar aquestes relacions i predir possibles resultats en futures situacions. Ara bé, és important tenir en compte que la precisió dels resultats obtinguts depèn molt de com s'hagi realitzat l'anàlisi de dades, així com la qualitat de les suposicions.

En un principi pot semblar que l'anàlisi predictiu és el mateix que fer un pronòstic (que fa prediccions a un nivell macroscòpic), però es tracta d'una cosa completament diferent. Per exemple, un pronòstic pot predir quants gelats es vendran el mes que ve i l'anàlisi predictiu pot indicar quins individus és més probable que es mengin un gelat. Aquesta informació, si s'utilitza de forma correcta, suposa un canvi radical en el joc, ja que permet orientar els esforços per ser més precisos a l'hora d'aconseguir objectius.

Per realitzar un anàlisi predictiu és imprescindible disposar d'una quantitat considerable de dades, tant actuals com passades, per poder establir patrons de comportament i així induir coneixement. Per exemple, en el cas comentat en el paràgraf anterior, si es creuen dades sobre la temperatura registrada, l'època de l'any i si és cap de setmana o festiu, es pot deduir quin tipus de perfil de persona menjarà un gelat. Aquest procés es realitza gràcies a l'aprenentatge computacional. Els ordinadors poden "aprendre" de manera autònoma i d'aquesta manera desenvolupar un nou coneixement, tot això a partir de proporcionar-lis el més potent recurs natural de la societat moderna: les dades.

4.2. Models aplicables a l'anàlisi predictiu

Generalment s'utilitza el terme anàlisi predictiu quan en realitat s'està parlant de modelat predictiu, que realitza qualificacions mitjançant models predictius i pronòstics.

Un model predictiu és un mecanisme que prediu el comportament d'un individu. Utilitza les característiques de l'individu com entrada i proporciona una qualificació predictiva com a sortida. Quant més elevada sigui la qualificació, més alta és la probabilitat que l'individu dugui a terme el comportament predit.

La qualificació obtinguda per qualsevol model predictiu ha de ser tinguda en compte amb especial cura i pot requerir que es compari amb un altre model o que es faci un anàlisi addicional a l'hora d'aplicar-la a un individu concret. Les qualificacions parlen de tendències i possibilitats en un grup suficientment gran, però no garanteix que la predicció es compleixi en cada cas individual.

Això es pot entendre millor amb un exemple: en una entitat financera el model aplicat a l'anàlisi de riscos preveu per un determinat perfil de clients que l'incompliment en un rebut mensual de la targeta de crèdit quadruplica la possibilitat que no pagui algun altre rebut durant aquell any. Tot i així, un anàlisi que inclogui noves variables pot concloure que

l'impagament d'aquell mes ha sigut produït a causa d'una despesa extra no previsible (com una avaria en el cotxe) i canviar la predicció de possibilitat d'un altre impagament durant l'any.

El tipus d'anàlisi que permeten els models predictius valora la relació existent entre cents d'elements per aïllar les dades que informen sobre un fet, guiant a la presa de decisions per un camí segur.

4.2.1. Aprenentatge supervisat i no supervisat

Cal diferenciar si el problema que se'ns presenta és un cas d'aprenentatge supervisat o no supervisat. [19]

En l'aprenentatge supervisat, els algoritmes treballen amb dades etiquetades (*labeled data*), intentant trobar una funció que, donades les variables d'entrada, els hi assigni l'etiqueta o valor de sortida adequada. L'algoritme s'entrena amb un històric de dades i així aprèn a assignar l'etiqueta o el valor de sortida adequat.

Alguns dels algoritmes més utilitzats en l'aprenentatge supervisat són:

- Arbres de decisió
- Classificació de Naïve Bayes
- Regressió per mínims quadrats
- Regressió Logística
- Support Vector Machines (SVM)
- Mètodes Ensembe (Conjunt de classificadors)

Per altra banda, l'aprenentatge no supervisat té lloc quan no es disposa de dades etiquetades per l'entrenament. Només es coneixen les dades d'entrada, però no existeixen dades de sortida que corresponguin a un determinat *input*. Per tant, només es pot descriure l'estructura de les dades per intentar trobar algun tipus d'organització que simplifiqui l'anàlisi. És per això que es diu que tenen un caràcter exploratori.

Alguns dels algoritmes més típics d'aprenentatge no supervisat són els algoritmes d'agrupament (*Clustering*). Consisteix en un procediment d'agrupació d'una sèrie de vectors d'acord amb un criteri. Aquests criteris acostumen a ser la distància o la similitud.

Altres algoritmes molt habituals en l'aprenentatge no supervisat són:

- Anàlisis de components principals
- Descomposició en valors singulars
- Anàlisis de components independents

4.2.2. Tipus de predicció dins l'aprenentatge supervisat

S'identifiquen dos tipus d'anàlisi predictiu segons el tipus de variable objectiu. Per una banda estan els models de classificació que permeten predir la pertinència a una classe. Un exemple podria ser intentar classificar els clients més propensos a abandonar l'empresa. Els resultats són binaris (en forma de 0 o 1) amb els seu grau de probabilitat. Per altra banda, estan els models de regressió, que ens permeten predir un valor. Per exemple, quin és el benefici d'un determinat client en els pròxims mesos.

4.2.3. Validació

Una vegada s'ha creat el model és necessari comprovar que funciona de manera correcta. Es tracta de l'aspecte més important dels models predictius: la seva validació. Una manera molt estesa de fer-ho consisteix en dividir el conjunt de dades del que es disposa en dos. Per una banda, es té un conjunt de dades sobre el qual es desenvolupa el model. Aquest acostuma a consistir en dues terceres parts de la mostra i s'anomena conjunt d'entrenament (*training test*). Per altra banda, la tercera part sobrant s'utilitza per validar el model i s'anomena conjunt de test (*test set*).

4.3. Tècniques aplicables a l'anàlisi predictiu

Els enfocaments i tècniques utilitzats per realitzar l'anàlisi predictiu poden agrupar-se d'una manera molt general en tècniques de regressió i tècniques d'aprenentatge computacional.

4.3.1. Tècniques de regressió

Els models de regressió són el pilar fonamental de l'analítica predictiva. Es basa en l'establiment d'una equació matemàtica com a model per representar les interaccions entre les diferents variables en consideració. Depenent de la situació, hi ha una gran varietat de models que es poden aplicar durant la realització de l'anàlisi predictiu. [14]

- Regressió lineal

  El model de regressió lineal analitza la relació existent entre la variable dependent o de resposta i un conjunt de variables independents o predictores. Aquest relació s'expressa com una equació que prediu la variable de resposta com una funció lineal dels paràmetres. Aquests paràmetres es calculen per tal que la mesura d'ajust sigui òptima. Gran part de l'esforç per l'adaptació del model es centra en minimitzar l'error, així com assegurar-se que s'està distribuint de forma aleatòria respecte a la predicció del model.

A la Figura 4 es pot veure un exemple de regressió lineal simple, on certs valors de la variable dependent queden per sobre de la funció lineal general i altres per sota.

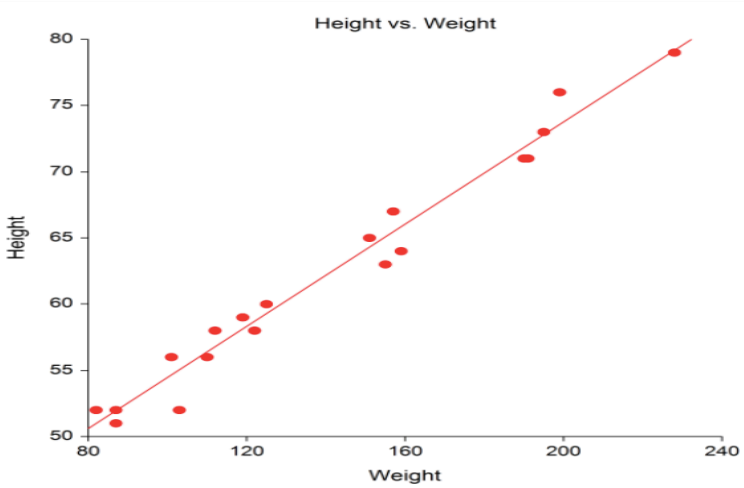

Figura 4. Exemple de regressió lineal [23]

- Anàlisi de supervivència o duració

  Es tracta d'un anàlisi del temps transcorregut fins un determinat esdeveniment. La censura i la no normalitat, que són característiques de les dades de supervivència, generen dificultat a l'hora d'intentar analitzar les dades fent servir models estadístics convencionals com la regressió lineal múltiple.

  Un concepte important en l'anàlisi de supervivència és la taxa de risc, definida com la probabilitat de que un esdeveniment passi en el temps *t*, cosa que implica sobreviure fins el temps *t*. Un altre concepte relacionat és la funció de supervivència, que pot definir-se com la probabilitat de sobreviure al temps *t*.

- Arbres de classificació i regressió

  Els arbres de classificació i regressió (*Classification And Regression Trees*, CART) estan formats per una col·lecció de regles basades en les variables del model. Aquestes regles es seleccionen per obtenir la millor divisió possible tenint en compte tant les variables categòriques com les contínues. Una vegada que es divideix un node en dos, s'aplica el mateix procediment als nodes 'secundaris', és a dir que es tracta d'un procediment recursiu. La divisió s'atura quan es detecta que ja no es pot millorar el resultat o quan es compleixen algunes regles de parada preestablertes.

- Corbes de regressió adaptativa multivariable

  Les corbes de regressió adaptativa multivariable (*Multivariate Adaptative Regression Splines*, MARS) són una tècnica no paramètrica que construeix models flexibles a l'ajustar regressions lineals per peces. Un concepte important és el node, que és on un model de regressió local dona pas a un altre i, per tant, és el punt d'intersecció entre dues corbes.

Cal destacar que aquestes no són les úniques tècniques de regressió i que n'existeixen d'altres com les d'elecció discreta, de regressió logística, de regressió logística multinomial, els models probit o els models de sèries temporals.

4.3.2. Tècniques d'aprenentatge computacional

L'aprenentatge computacional es va emprar originalment per desenvolupar tècniques que permetessin als ordinadors aprendre. Avui en dia, aquestes tècniques, a l'incloure una sèrie de mètodes estadístics avançats per la regressió i la classificació, tenen aplicacions en una àmplia varietat de camps. [17]

- Xarxes neuronals

  Les xarxes neuronals són tècniques de modelat no lineal sofisticades que són capaces de modelar funcions complexes. Poden aplicar-se a problemes de predicció, classificació o control en un ampli espectre de camps.

  Les xarxes neuronals s'utilitzen quan no es coneix la naturalesa exacte de la relació entre els valors d'entrada i de sortida. Una característica clau de les xarxes neuronals és que aprenen de la relació entre els valors d'entrada i sortida a través de l'entrenament. Existeixen tres tipus d'entrenament en les xarxes neuronals: l'aprenentatge per esforç, el supervisat i el no supervisat, sent el supervisat el més comú.

- Màquines de vectors de suport

  Les màquines de vectors de suport (SVM) s'utilitzen per detectar i explotar patrons complexos de dades a partir d'agrupar-les, ordenar-les i classificar-les. Són màquines d'aprenentatge que s'utilitzen per realitzar classificacions binaries i estimacions de regressió. Usualment utilitzen mètodes per aplicar tècniques de classificació lineal a problemes de classificació no lineal. Hi ha diferents tipus de SVM: lineal, polinomial, sigmoide, etc.

- Naïve Bayes

  El classificador bayesià ingenu es basa en la regla de la probabilitat condicional de Bayes, que s'utilitza per la tasca de classificació. El classificador bayesià assumeix que les variables predictores són estadísticament independents, cosa que fa que sigui una eina de classificació fàcil d'interpretar. És especialment útil quan el número de prediccions és molt alt.

- K-veïns més propers

  L'algoritme de veïns més pròxims k-NN (Nearest Neighbor) pertany a la classe de mètodes estadístics de reconeixement de patrons. El mètode no imposa a priori ninguna suposició sobre la distribució de la mostra de dades.

  A diferència dels altres mètodes, aquest és asimptòticament convergent, és a dir, a mesura que la mida del conjunt d'entrenament augmenta, si les observacions són independents i idènticament distribuïdes, la classe que s'ha predit convergirà a l'assignació de la classe que minimitzi l'error de la classificació errònia.

Igual que en l'apartat anterior, aquestes no són les úniques tècniques d'aprenentatge computacional. Existeixen altres com la funció de base radial, el perceptró multicapa o el modelat predictiu geoespacial.

5. Descripció de la instal·lació fotovoltaica de l'ETSEIB

La instal·lació fotovoltaica de l'ETSEIB va néixer a partir del projecte BISOL, una iniciativa que consisteix en l'electrificació de les taules d'estudi de la biblioteca de l'ETSEIB amb energia renovable per tal de carregar tauletes, portàtils o altres aparells electrònics. [37]

La instal·lació ha estat dissenyada per una potència de 4kW i té una autonomia d'un dia. Està composta per 16 panells solars fotovoltaics situats a la coberta de la biblioteca de l'ETSEIB, un inversor híbrid amb un regulador de càrrega integrat, unes bateries i un sistema d'adquisició i monitorització de dades. L'esquema bàsic de la instal·lació es mostra a la Figura 5.

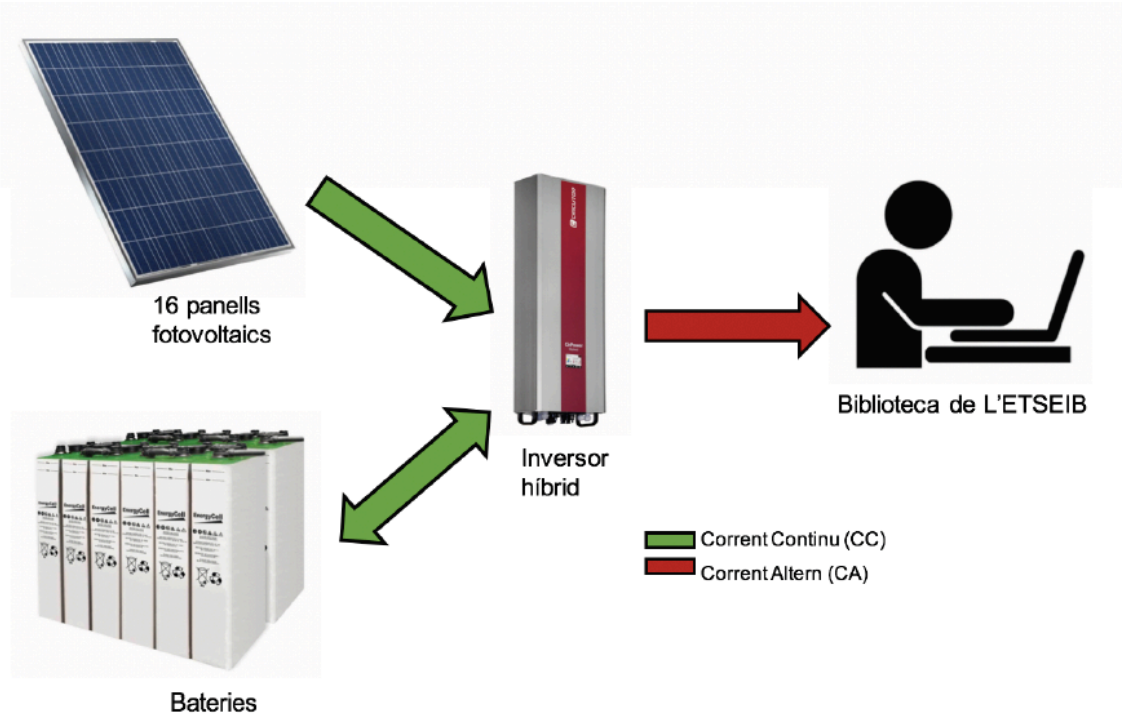

Figura 5. Esquema bàsic de la instal·lació autònoma de l'ETSEIB [29]

Els 16 panells solars fotovoltaics són de la marca Atersa i estan compostos de cèl·lules policristal·lines de 156x156 mm de mida. Per connectar els panells entre ells es fan servir connectors de la marca TYCO. Es va decidir distribuir-los en 4 files, de 4 panells cadascuna, i inclinar-los 60 graus per garantir la màxima potència en els mesos de menys insolació. Es calcula que la producció energètica en el millor mes de l'any (juliol) és de 141,9 kWh/m^2 i en el pitjor mes de l'any (febrer) de 110,2 kWh/m^2. La producció total és d'aproximadament 1514,1 kWh/m^2. [4]

De les plaques solars surt un cablejat que porta a una caixa de proteccions de 24V i després a l'inversor de càrrega híbrid solar, que és de la marca Circutor. Aquest està especialment pensat per instal·lacions fotovoltaiques autònomes i incorpora un regulador de càrrega.

Per tal d'allargar al màxim l'estat de la bateria s'utilitza el Cirpower Hybrid 4k-48V, que es pot posar en mode "aïllat" per tal de desconnectar totalment l'equip de la xarxa. Aquesta opció s'aplica en funció de l'estat de càrrega de les bateries (SOC) [7]. Més concretament, si es parteix d'un SOC inferior al mínim prèviament determinat, l'equip no connectarà les línies fins que aquest superi un valor de SOC recomanable prèviament determinat. D'aquesta manera, els panells solars es fan servir en primer lloc per subministrar energia als endolls de la biblioteca, en el cas que hi hagi excés de generació per carregar la bateria i, si aquesta no permet més càrrega perquè el SOC està al 100%, es reduirà la generació.

De l'inversor surten dos cablejats: un de corrent continu (CC) cap a les bateries i un altre de corrent altern (CA) cap als endolls.

El flux de corrent continu (CC) va directament cap a les bateries, que són de plom, estacionàries i de baix manteniment. Concretament, són el model TOPzS de la marca Vesna Solar. Proporcionen una capacitat de 21,8 kWh i són ideals per instal·lacions autònomes, ja que suporten càrregues irregulars i estan dotades d'una descàrrega baixa. [36]

Quant al flux de corrent altern (CA), primer es dirigeix a una caixa de proteccions de 24V i sortint d'aquesta va a una segona caixa de sobretensions de 36V. En aquesta última caixa, a través de 3 contactors es divideix el flux en 3 línies, cadascuna de les quals alimenta 20 endolls mitjançant els canals UNEX73.

5.1. Dispositius d'adquisició i monitorització de dades de la instal·lació

A la instal·lació solar fotovoltaica de l'ETSEIB s'hi poden trobar diferents dispositius per tal d'adquirir i monitoritzar les dades del sistema, els quals queden descrits en aquest apartat. Els elements, de la marca Circutor, són els següents: tres analitzadors de xarxa CVM-1D, un inversor híbrid model CirPower 4k i un gestor energètic EDS Deluxe.

L'esquema bàsic del sistema d'adquisició i monitorització de dades es mostra a la Figura 6.

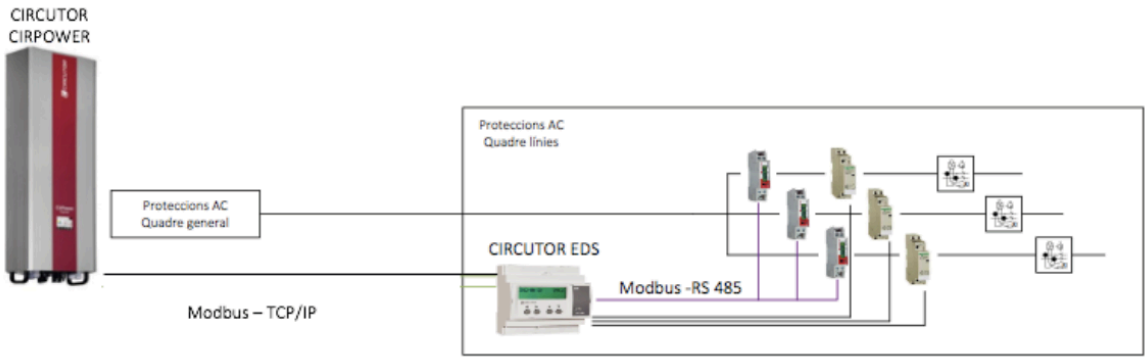

Figura 6. Esquema del muntatge dels dispositius d'adquisició de dades de la instal·lació [29]

- CVM-1D

  Es tracta d'una analitzador de xarxa i s'encarrega de mesurar, calcular i visualitzar els principals paràmetres elèctrics de la xarxa. N'hi ha un per cada línia de consum.

Figura 7. Analitzador de xarxa CVM-1D de Circutor [10]

- CirPower Hybrid 4k-48

  És un inversor híbrid solar dissenyat per instal·lacions fotovoltaiques aïllades. Apart de realitzar la funció de inversor (transformar el corrent continu en altern), també regula el flux energètic entre els panells solars, la càrrega i les bateries. A més a més, disposa d'un servidor web amb dades i gràfics que permet monitoritzar les principals variables del sistema en tot moment.

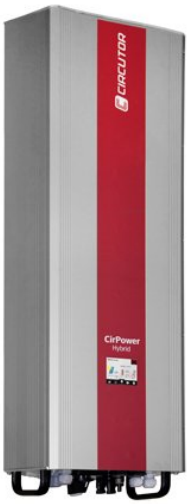

Figura 8. Inversor híbrid solar CirPower Hybrid 4k-48 de Circutor [7]

- EDS (Efficiency Data Server)

  Es tracta d'un gestor d'eficiència energètica amb el qual es poden visualitzar, controlar, gestionar i emmagatzemar les seves pròpies variables i també les d'aquells dispositius que hi estan connectats. L'usuari pot veure a temps real el valor de les entrades i sortides dels diferents dispositius mitjançant un servidor web accessible des d'un explorador d'internet convencional. Això permet centralitzar les dades recollides i accedir a les d'aquells dispositius que no disposen d'una interfície pròpia, com en el cas dels analitzadors de xarxa CVM-1D.

Figura 9. Gestor d'eficiència energètica EDS de Circutor [8]

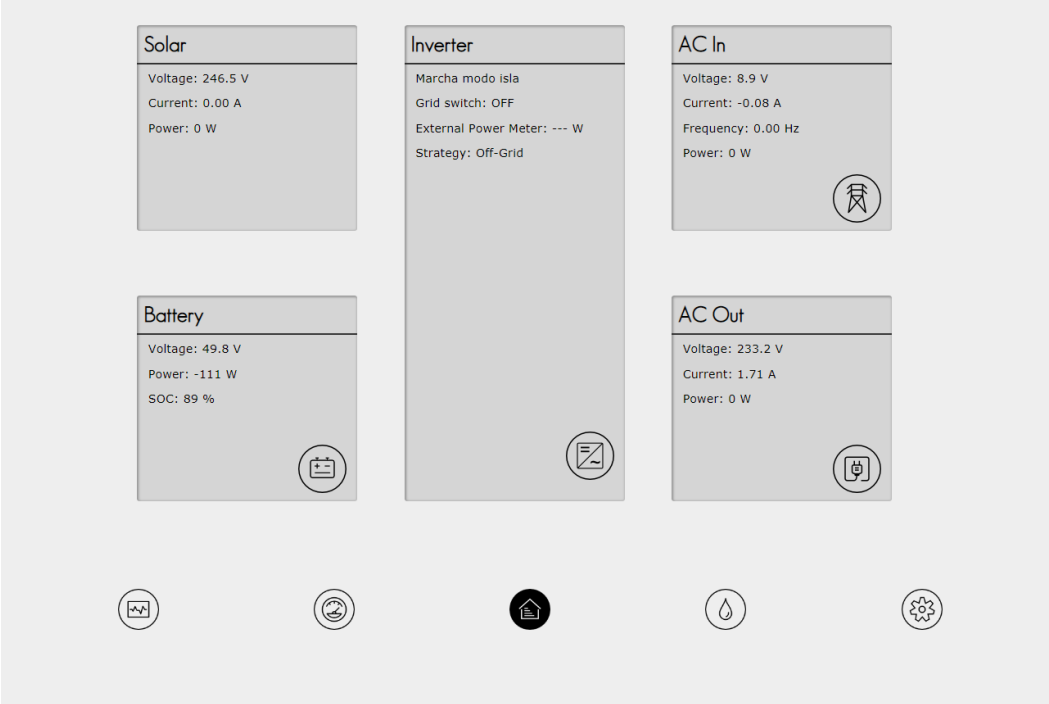

Figura 10. Servidor web del gestor d'eficiència energètica EDS de la instal·lació fotovoltaica de l'ETSEIB [12]

5.2. Software PowerStudio SCADA

Es tracta d'un software de la marca Circutor que permet visualitzar a temps real els paràmetres dels diferents elements que componen una instal·lació fotovoltaica. És capaç d'integrar tota mena d'equips de la marca Circutor. L'usuari pot crear pantalles, informes, gràfics i taules personalitzades amb la informació desitjada, així com també programar esdeveniments. [9]

Es divideix en tres mòduls:

- L'editor (PowerStudio SCADA editor): s'encarrega de gestionar les aplicacions i permet a l'usuari crear pantalles, gràfics i informes.
- El motor (PSEngineManager): executa l'aplicació creada per l'editor i es comunica amb els dispositius de la instal·lació.
- El client (PowerStudio Client): permet a l'usuari consultar pantalles, informes, gràfics o qualsevol informació creada prèviament amb l'editor.

En el cas de l'ETSEIB, els dispositius CVM-1D i el CirPower Hybrid 4k-48 es connecten al gestor d'eficiència energètica EDS, que és el motor del software. Per tal d'aconseguir les

dades necessàries per fer la predicció, es fa ús del mòdul client des del servidor web que ofereix SCADA. Per entrar-hi cal estar connectat a la xarxa de l'ETSEIB i escriure la direcció IP [11] al buscador com a HTTP.

Tal i com es pot veure a la Figura 11, la pantalla principal mostra informació en gràfics sobre l'estat dels dispositius i els esdeveniments. Clicant a la icona de la part superior esquerra de la pantalla hi apareix un menú, el qual es pot veure a la Figura 12.

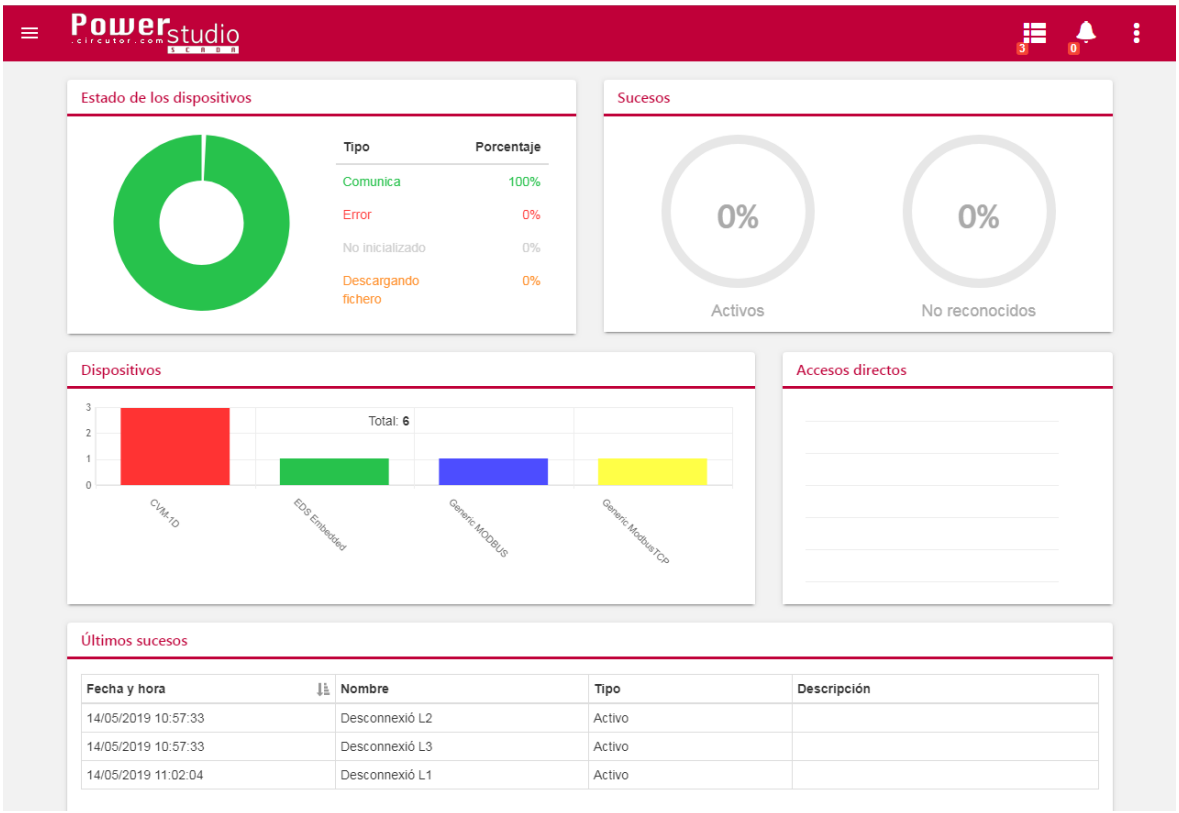

Figura 11. Pantalla principal del servidor web de SCADA aplicat a la biblioteca de l'ETSEIB [11]

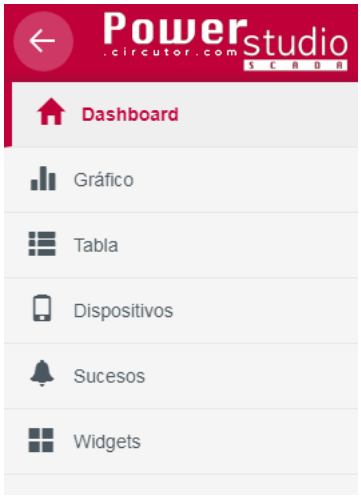

Figura 12. Menú desplegable del servidor web de SCADA [11]

En aquest projecte s'utilitza l'opció taula, ja que, a diferència de les altres, permet visualitzar i descarregar un conjunt de dades de dimensió considerable en un format compatible amb programes externs d'anàlisi. En clicar sobre l'opció apareix un menú que permet escollir entre els diferents dispositius que adquireixen dades de la instal·lació, així com esdeveniments del sistema. Aquest es mostra a la Figura 13.

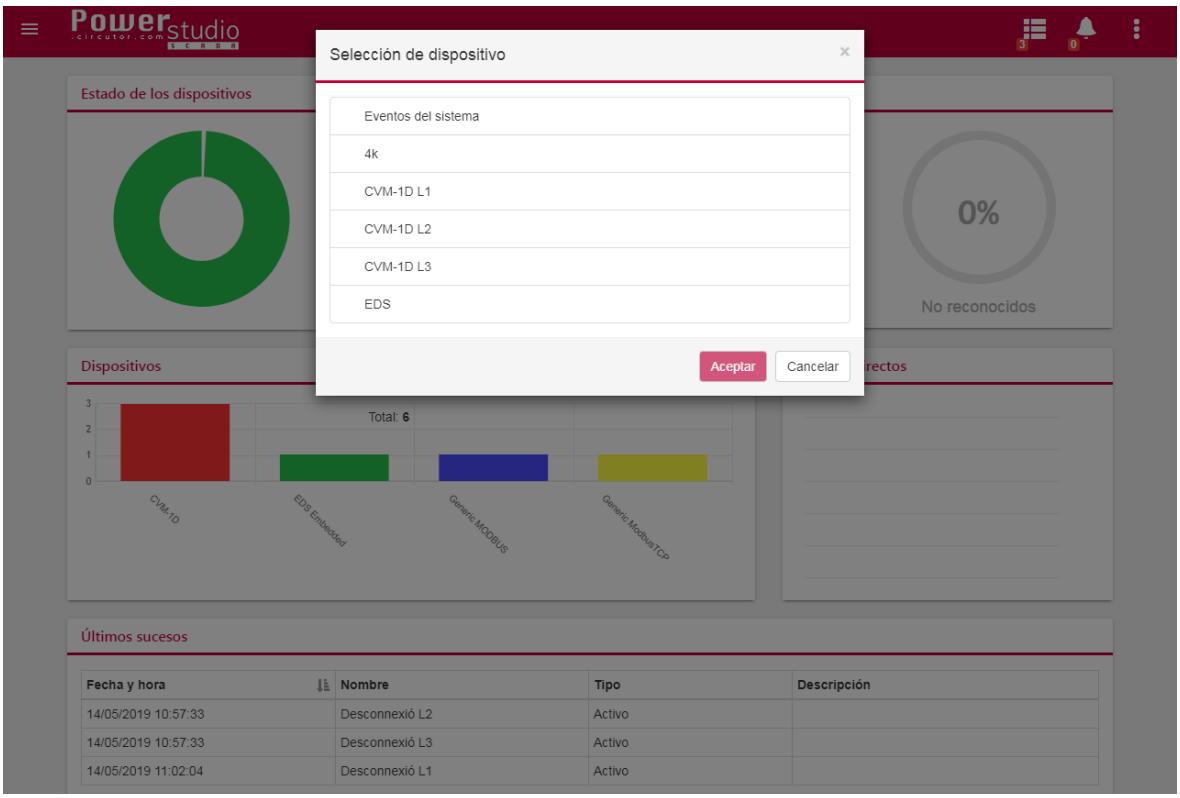

Figura 13. Pantalla emergent del servidor web SCADA amb les diferents opcions de dispositius d'adquisició de dades a la instal·lació fotovoltaica de l'ETSEIB [11]

Un cop triat el dispositiu, apareix un altre menú en el que es poden triar les variables que es volen visualitzar, que són diferents per a cada dispositiu. Es mostra el cas d'una de les línies de la instal·lació a la Figura 14.

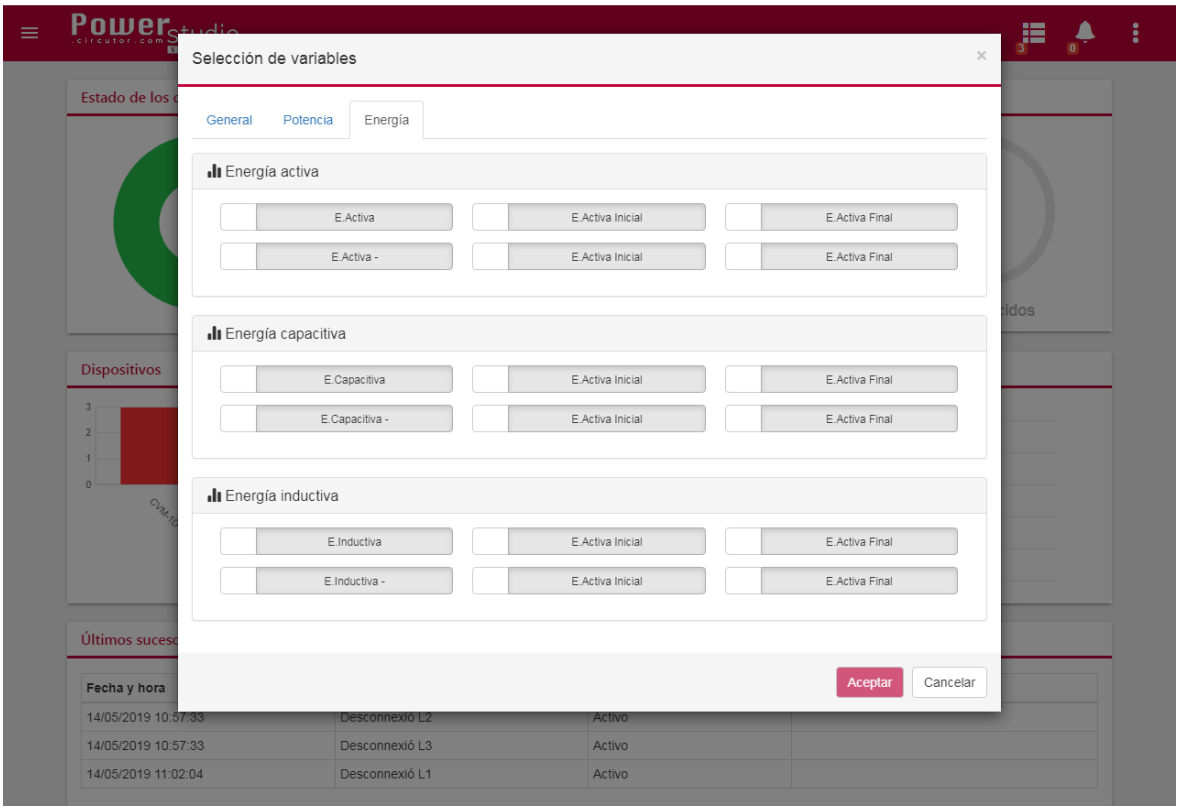

Figura 14. Menú emergent del servidor web SCADA amb les diferents opcions de variables enregistrades en una de les línies de la instal·lació fotovoltaica de l'ETSEIB [11]

Més endavant, també es poden triar la freqüència de la generació de les dades i el període de temps en el que es volen mostrar les dades seleccionades, tal i com es veu a la Figura 15.

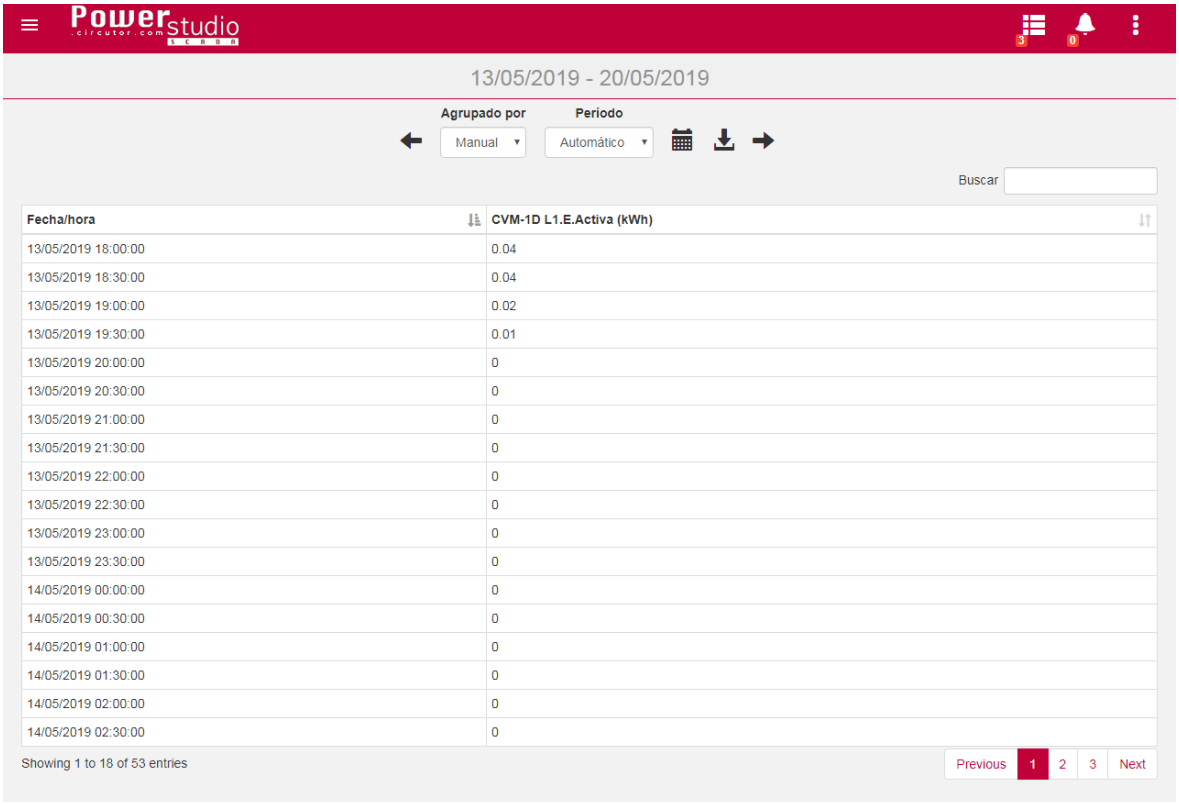

Figura 15. Pantalla del servidor web SCADA amb alguns valors d'energia activa enregistrats a una de les línies de la instal·lació fotovoltaica de l'ETSEIB [11]

Finalment, es poden descarregar les dades en un format csv. Es tracta d'un format que, per la seva senzillesa, és utilitzat per importar i exportar gran quantitat de dades.

6. Metodologia per fer prediccions

En aquest apartat s'explica la metodologia seguida per tal de fer un anàlisi predictiu. L'estructura principal consisteix en la descàrrega, anàlisi i tractament de les dades, la comparació de la qualitat que ens aporten diferents algoritmes, l'elecció d'un d'ells i l'optimització d'aquest. [38]

L'esquema és el mostrat a la Figura 16.

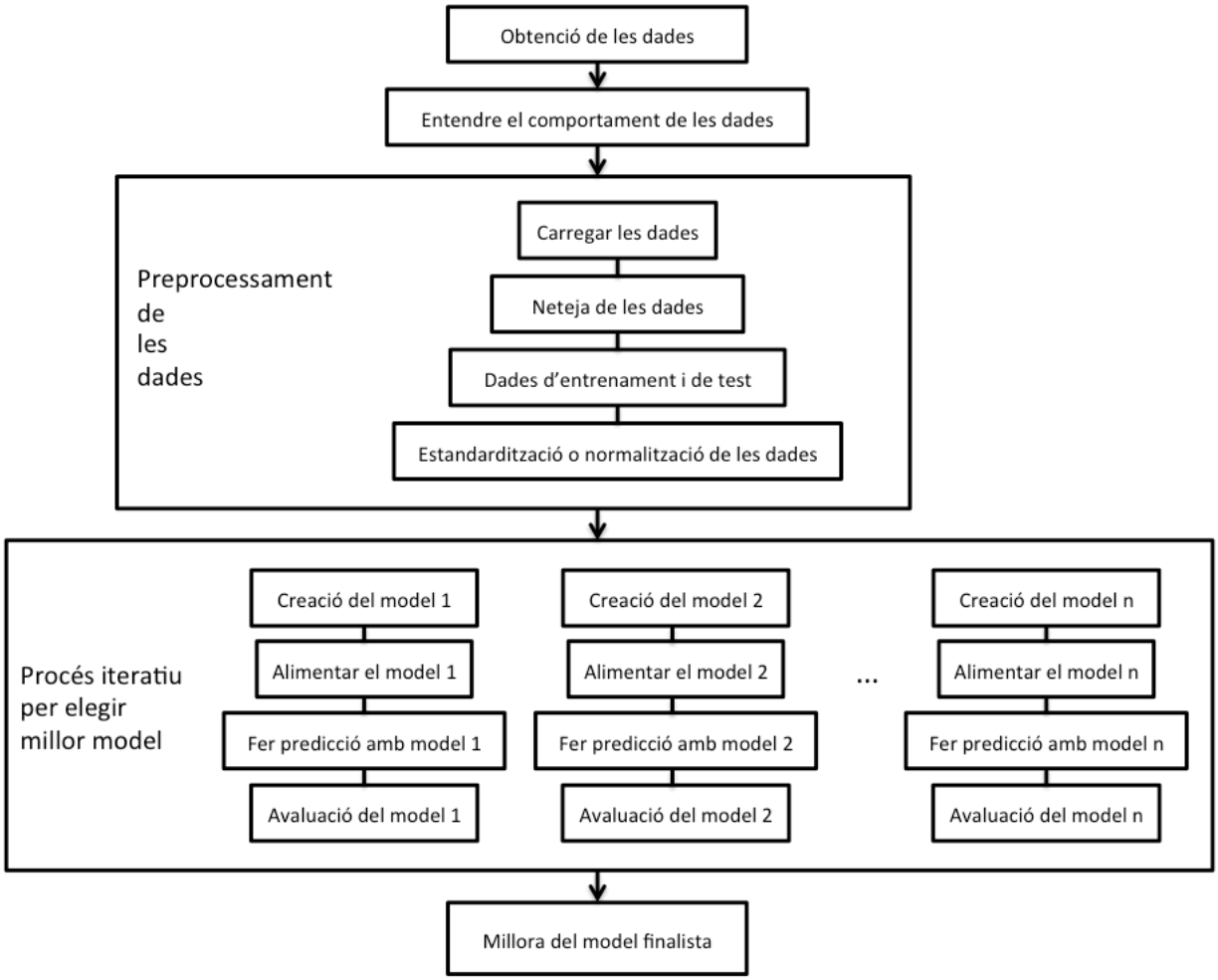

Figura 16. Esquema de la metodologia emprada per fer prediccions

Cal tenir en compte que durant aquest procés s'utilitzen funcions que empren algun paràmetre aleatori. Aquest paràmetre, que varia cada cop que es crida la funció en cas que no se li digui el contrari, fa que els resultats d'una mateixa línia de codi vagin canviant cada cop que es crida aquesta. Per tal d'evitar que això passi i poder arribar als mateixos resultats cada cop que es crida una funció concreta, es fa servir durant tot l'estudi la mateixa llavor

aleatòria (*random seed*), la qual s'ha d'indicar en cada funció que tingui aquest caràcter aleatori.

6.1. Obtenció de dades

Cal aconseguir les dades necessàries per fer la predicció. Tant aquelles que es volen predir com aquelles que es creu que poden mantenir una relació amb el valor de la predicció.

6.2. Entendre el comportament de les dades

Abans de començar a tractar les dades, és important fer un estudi previ per entendre:

- Número de dades o mostres
- Tipus de variable per cada atribut
- Resum de la descripció estadística: mitjana, variació, valor mínim i màxim, percentils, ...
- Correlació entre atributs
- ...

6.3. Preprocessament de les dades (*Preprocessing*)

El procés consisteix en convertir el conjunt de dades brutes en dades netes i rellevants. Es tracta d'un pas essencial abans d'alimentar l'algoritme amb les dades. Hi ha tres passos principals a seguir:

6.3.1. Càrrega de les dades

Es necessiten les dades de forma numèrica i en matrius.

6.3.2. Neteja de les dades

Aquest pas consisteix en validar:

- Dades buides o nul·les en les columnes
- Valors molt allunyats de la tendència que es puguin considerar com erronis

6.3.3. Dades d'entrenament i de test

El següent pas és separar les dades entre el conjunt d'entrenament i el de test (*Train Test Split*). Les dades d'entrenament serveixen per crear i millorar el model predictiu i, un cop creat, s'utilitzen les dades de test per jutjar la qualitat del model. No hi ha un percentatge que marqui la relació ideal entre tos dos conjunts, però s'acostuma a fer un 67% per les dades d'entrenament i un 33% per les de test.

Per altra banda, les dades d'entrenament també es poden dividir entre varis grups per tal d'aplicar més endavant el mètode de validació creuada (*Cross-Validation*). Aquest mètode consisteix en entrenar i testejar el model varis cops per tal d'obtenir més d'un resultat de la qualitat de la predicció. D'aquesta manera, la comparació entre models, que s'explica més endavant, és més fiable.

6.3.4. Estandardització o normalització de les dades

Existeixen alguns algoritmes en el que les dades milloren significativament el seu comportament quan es presenten d'una certa manera. Les principals dues tècniques utilitzades són:

- Normalització: s'aplica quan els atributs són d'ordres de magnitud molt diferents i consisteix en escalar les dades perquè es mantinguin dins d'un interval entre 0 i 1.

- Estandardització: s'aplica quan les dades presenten una distribució normal i les transforma per tal que tinguin una mitjana de 0 i una variància de 1.

En el cas que sigui convenient aplicar alguna d'aquestes dues tècniques, la metodologia seguida consisteix en normalitzar o estandarditzar les dades d'entrenament i seguidament calcular la mitjana (en el cas d'estar normalitzant les dades) o la mitjana i la variància (en el cas d'estar aplicant estandardització). Aquest valors serveixen per poder normalitzar o estandarditzar de la mateixa manera les dades de test i, un cop acabat el model, per entrar els atributs de manera correcta quan es vulgui fer una predicció.

6.4. Elecció del model

Després de fer totes les transformacions necessàries a la base de dades, és hora de triar el model més adient o l'algoritme que millor representa les dades que s'estan treballant. En un principi no es pot saber quin model oferirà a les dades del problema una previsió més acurada. És per això que és necessari provar-los i comparar la qualitat de les prediccions que facin. Això implica que els següents passos s'han de repetir per cadascun dels models:

6.4.1. Creació del model

Primer de tot s'ha de crear el model a partir d'algun algoritme d'aprenentatge automàtic. Cal tenir en compte el tipus de predicció que s'està fent per tal de veure quins algoritmes es poden valorar com a opció. Concretament, cal fixar-se especialment en si l'aprenentatge és supervisat o no supervisat i en el tipus de predicció que s'està realitzant: classificació, regressió o *clustering.*

6.4.2. Alimentar el model

Seguidament s'alimenta el model amb les dades d'entrenament (tant les variables independents com la que s'està intentant predir) per tal que aprengui com comportar-se amb aquest tipus de dades. Com més gran sigui la base de dades amb la que s'alimenti el model, més precises seran les previsions que faci.

6.4.3. Predicció

En aquest pas es demana al model que faci una predicció partint de les dades de test que fan referència a les variables independents, és a dir, sense donar-li el valor de la variable que s'està intentant predir.

6.4.4. Avaluació

Per tal d'avaluar la qualitat de la predicció que s'ha fet, existeixen diferents tècniques aplicades a l'aprenentatge automàtic. Aquestes variaran depenent del tipus de predicció que s'estigui fent, per exemple:

- Classificació
    - Matriu de confusió (*Confusion Matrix*)
    - Puntuació de la precisió (*Accuracy Score*)
- Regressió
    - Error absolut mig (*Mean Asolute Error*)
    - Error quadràtic mig (*Mean Squared Error*)
    - Coeficient de determinació R^2
- Algoritmes d'agrupament (*Clustering*)
    - Homogeneïtat

- V-measure
- Validació creuada (*Cross-validation*)

6.5. Millora del model

Un cop s'ha escollit el model que millor s'ajusta a les dades del projecte, es pot procedir a optimitzar-lo. Es tracta de buscar els paràmetres adequats pel nostre problema, és a dir, el valor que haurien de prendre els paràmetres del model escollit per tal que la predicció sigui el més precisa i exacte possible.

Existeixen eines que creen i validen models iterativament amb l'algoritme que s'especifiqui i utilitzant diferents valors de paràmetres per tal de informar quina opció dóna una més bona predicció. Els dos mètodes més coneguts són la cerca per quadrícula (*Grid Search*) i l'optimització de paràmetres aleatoris (*Randomized Parameter Optimization*).

A més a més, caldrà validar que totes les variables influencien en el resultat de la predicció, ja que més atributs no implica millors resultats. En el cas de trobar algun atribut que no influenciï pràcticament en el resultat de la predicció i que es decideixi eliminar-lo, caldrà tornar a realitzar els passos anteriors per confirmar que el model escollit és la millor opció.

7. Predicció de la demanda elèctrica de la biblioteca de l'ETSEIB

Tal i com s'ha explicat, el que es vol predir és la demanda de la instal·lació de plaques fotovoltaiques de l'ETSEIB. Aquestes s'utilitzen per alimentar els endolls de la biblioteca de l'escola i, per tant, el valor que defineix millor aquesta demanda és l'energia elèctrica consumida per carregar el que s'endolli.

Com que les dades de les quals es parteix també inclouen el resultat de la predicció, és un cas d'aprenentatge supervisat. A més a més, en estar-se intentant predir un valor, no una classe, es tracta d'una predicció de regressió.

També és necessari mencionar que els endolls només es poden utilitzar quan la biblioteca està oberta i, per tant, quan està tancada la demanda és zero. Es tracta d'un resultat conegut que no tindria sentit intentar predir. És per aquest motiu que es decideix crear un model que prediu la demanda quan la biblioteca està oberta. Aquesta informació es pot aconseguir amb el calendari de la biblioteca de l'ETSEIB, el qual indica els canvis d'horari en funció dels dies festius i el calendari acadèmic.

Per tal de començar cal definir la variable que defineix millor allò que es vol predir, així com aquelles variables que influeixen més en aquest valor. Per altra banda, també s'ha de pensar com es poden aconseguir aquestes dades.

7.1. Variable a predir

Tal i com s'ha dit, el valor que defineix millor la demanda de la instal·lació de plaques fotovoltaiques de l'ETSEIB és l'energia consumida per carregar el que s'endolli.

Aquesta informació es pot obtenir gràcies al Software PowerStudio SCADA de la marca Circutor, que permet visualitzar a temps real els paràmetres dels diferents elements que componen la instal·lació fotovoltaica [9]. Per altra banda, també emmagatzema tota la informació llegida cada 15 minuts i té un històric de dades dels darrers 6 mesos. Per la descàrrega d'aquestes dades s'utilitza el servidor web del client de PowerStudio SCADA [11]. Es pot trobar més informació a l'apartat 5.2. del present treball.

A més d'això, gràcies al treball de fi de grau realitzat al Gener del 2019 "Avaluació del funcionament d'un sistema fotovoltaic aïllat a l'ETSEIB" [29], el qual va analitzar les diferents variables de la instal·lació fotovoltaica i la seva evolució, es disposa de dades des del març del 2018.

Aquesta base de dades conté informació de diferents punts de la instal·lació fotovoltaica sobre variables com la tensió, el corrent, l'energia, la potència… Tot i així, per tal de fer aquest estudi només caldrà centrar-se en la variable que indica l'energia activa de cada línia, que és el valor mitjà de l'energia consumida en la respectiva línia durant el darrers 15 minuts.

Finalment, el valor que es predirà és la suma de totes aquestes energies consumides, el qual serà l'energia consumida total de la xarxa i, per tant, la demanda de la instal·lació.

7.2. Variables per fer la predicció

En un primer moment cal valorar totes aquelles variables que poden influenciar el valor final del que s'està intentant predir. Més endavant es confirmarà si realment totes les variables són rellevants.

- Hora
- Dia de la setmana
- Dia de l'any

  Es tracta d'una manera de quantificar l'època de l'any o, el que vindria a ser el mateix, el dia i mes. També serveix per veure si algunes dates especials (com podria ser el dia de Sant Jordi) fan canviar la demanda de la instal·lació fotovoltaica. L'objectiu és que cada dia de l'any tingui un número associat i aquest sempre sigui el mateix. És per això que cal tenir en compte que alguns anys tenen 366 dies. En aquests casos s'imposarà que el dia 28 i 29 de febrer tinguin el mateix número.

- Setmana de l'any

  Es tracta d'una altra manera de quantificar l'època de l'any de manera més general.

- Període o època del curs

  Aquesta variable diferencia els dies entre les següents èpoques:

  - Preparcial: Des del dia que comença el quadrimestre fins la setmana abans de parcials.
  - Parcials: Setmana en la que es realitzen les proves parcials a l'escola
  - Postparcials: Des de que acaben els exàmens parcials fins quan acaben les classes del quadrimestre.

- Finals: Des de que acaben les classes del quadrimestre fins que es realitza l'últim examen final a l'escola.
- Reavaluacions: Des de que acaben els exàmens finals del quadrimestre de primavera fins que es realitza l'última prova de reavaluació.
- Altres: Aquesta categoria serveix per aquells dies que no entren dins cap de les altres descripcions. Concretament, es tracta de les dues setmanes de festa entre els quadrimestres de tardor i primavera, i l'època entre les reavaluacions i l'inici del quadrimestre de tardor.

- Horari que fa la biblioteca aquell dia

  Depenent del dia, la biblioteca obra a les 8:30h, 9h o 10:30h i tanca a les 14h, 20h, 20:30h o 22h. Podria ser que, per exemple, si la biblioteca obra de 9h a 14h, menys alumnes decideixin anar expressament a la biblioteca.

- Número de setmanes fins exàmens

  Per tal de quantificar la proximitat dels exàmens.

Totes aquestes dades s'obtindran gràcies al calendari acadèmic de l'escola [15] i el de la biblioteca [5].

Hi ha altres variables que es creuen que podrien ajudar a predir el valor de la demanda però que, donat que només es disposa de dades de menys de tres quadrimestres, no es pot demostrar que el seu valor influenciï realment en la demanda. És per aquest motiu que no s'afegeixen al model. Aquestes variables són:

- Número d'estudiant matriculats aquell quadrimestre a l'ETSEIB
- Variable categòrica que mostri si aquell any les proves parcials del segon quadrimestre cauen abans o després de setmana santa.

7.3. Eines utilitzades

Les principals eines utilitzades en aquest treball són Python, Pandas, Scikit-learn i Matplotlib.

- Python

  Es tracta d'un llenguatge de programació interpretat. És un llenguatge que simplifica molt la programació i la fa més àgil. Té una estructura ordenada i neta que permet que els codis siguin fàcils de interpretar. Es tracta d'un producte de codi obert. [26]

Encara que va ser creat com un llenguatge de programació d'ús general, conté una àmplia sèrie de llibreries i entorns de desenvolupament per cadascuna de les fases del procés de la ciència de dades.

- Pandas

  Pandas és una llibreria de Python destinada a l'anàlisi de dades, que proporciona unes estructures de dades flexibles i que permet treballar amb elles de forma molt eficient. En concret, ofereix estructures de dades i operacions per manipular taules numèriques i sèries temporals. Està escrita com una extensió de NumPy i és un software lliure. [24]

- Scikit-learn

  Es tracta d'una biblioteca pel llenguatge Python d'aplicació a l'aprenentatge automàtic. Inclou varis algoritmes de classificació, regressió i anàlisi de grups. Està dissenyada per interpretar les biblioteques numèriques NumPy i SciPy. [30]

- Matplotlib

  Matplotlib és una biblioteca de Python amb extensió matemàtica NumPy. Serveix per la generació de gràfics en 2D de qualitat en una varietat de formats i entorns interactius. Permet crear histogrames, gràfics de barres, diagrames de dispersió, diagrames de caixes, etc. en poques línies de codi. [20]

8. Resultats

Per la correcta interpretació de tots els subapartats cal tenir en consideració que l'ordinador utilitzat per arribar a resultats i executar les funcions és un portàtil MacBook, el qual té un processador de 2,4 GHz Intel Core 2 Duo i una memòria de 6 GB 1067 MHz DDR3.

8.1. Aconseguir les dades

Tal i com s'ha explicat, per aconseguir la base de dades són necessaris el servidor web del client de PowerStudio SCADA, el calendari de la biblioteca de l'ETSEIB i el calendari acadèmic de l'escola.

Cal tenir en compte que, en un primer moment, la informació de la que es disposa no es troba en un format tractable i cal especificar com es desitja tenir cada variable.

L'energia consumida, el valor de la qual s'està intentant predir, es dóna en KWh i dos decimals. Representa la suma de l'energia consumida entre les tres línies de la xarxa de la biblioteca de l'ETSEIB.

Pel que fa a les variables o atributs per fer la predicció, es presenten de la següent forma:

- Hora: variable quantitativa que es donarà en la unitat d'hores. Per exemple, si són les 19:15h, aquesta variable prendrà el valor 19,25h.

- Dia de la setmana: variable categòrica que indica si es tracta de dilluns, dimarts, dimecres, dijous, divendres, dissabte o diumenge. Seguint aquest ordre, prendrà el valor 1 quan sigui dilluns, 2 quan sigui dimarts… fins a valer 7 quan sigui diumenge.

- Dia de l'any: variable que indica el dia de l'any. Aquest número s'assignarà de forma ordenada, valent 1 el primer dia de gener i 365 el 31 de desembre. Tal i com s'ha explicat, en els anys de traspàs s'assignarà el mateix número pels dies 28 i 29 de febrer.

- Setmana de l'any: variable que indica el número de setmana de l'any.

- Període o època del curs: aquesta variable prendrà els valors següents:

Període del curs	**Valor de la variable**
Preparcials	1
Parcials	2
Postparcials	3
Finals	4
Reavaluacions	5
Altres	6

Taula 2. Valors que pren la variable del període del curs

- Horari que fa aquell dia la biblioteca: aquesta variable prendrà els valors següents:

Horari de la biblioteca	**Valor de la variables**
8:30h – 20:30h	1
8:30h – 14h	2
9h – 22h	3
8:30h – 22h	4
8:30h – 17h	5
9h – 14h	6
10:30h – 20h	7

Taula 3. Valors que pren la variable de l'horari de la biblioteca

- Número de setmanes fins exàmens: aquesta variable prendrà el valor 0 en els períodes de parcials, finals i reavaluacions. En les èpoques de preparcials i postparcials valdrà el número de setmanes que queden fins exàmens. Per exemple, si els parcials són la setmana que ve, la variable valdrà 1. En el cas de 'Altres', aquesta variable no té sentit i no s'ha de tenir en compte, ja que no hi ha cap quadrimestre en curs.

El fitxer amb les dades en aquest format s'aconsegueix programant amb Python i fent ús de la llibreria datetime [27]. Concretament, les funcions utilitzades, el codi de les quals es pot trobar a l'Annex A, són les mostrades a l'esquema de la Figura 17.

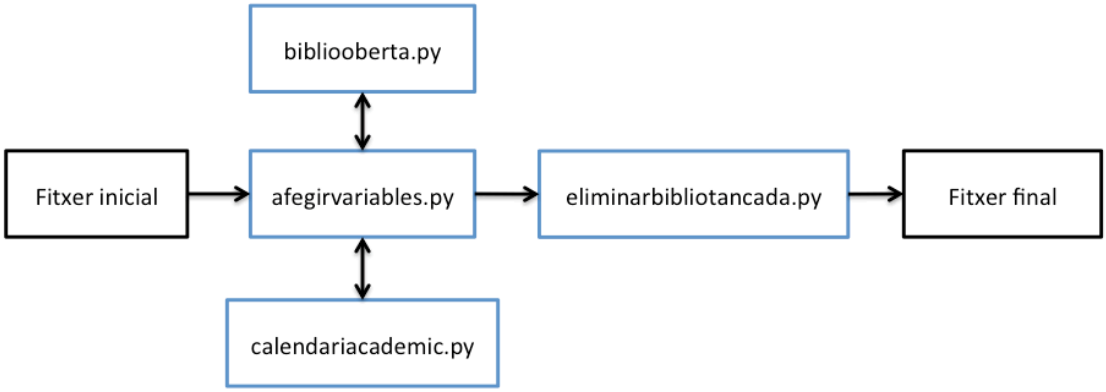

Figura 17. Esquema de les funcions necessàries per preparar les dades

El fitxer inicial consisteix en quatre columnes. La primera indica la data i l'hora a la que es va prendre la mostra, i les altres tres indiquen l'energia activa a cada línia en aquell moment. El format de cadascuna es pot veure a la Figura 18.

Fecha/hora	CVM-1D L1,E,Activa (kWh)	CVM-1D L2,E,Activa (kWh)	CVM-1D L3,E,Activa (kWh)
20/3/18 19:00			
20/3/18 19:15	0	0,02	0,01
20/3/18 19:30	0,01	0,02	0
20/3/18 19:45	0	0,01	0
20/3/18 20:00	0	0,01	0
20/3/18 20:15	0	0	0
20/3/18 20:30	0	0	0
20/3/18 20:45	0	0	0
20/3/18 21:00	0	0	0
20/3/18 21:15	0	0	0
20/3/18 21:30	0	0	0
20/3/18 21:45	0	0	0
20/3/18 22:00	0	0	0
20/3/18 22:15	0	0	0
20/3/18 22:30	0	0	0

Figura 18. Primeres files del fitxer inicial de la preparació de les dades

A la Figura 19 es mostren les primeres files del fitxer final. Es pot comprovar com amb aquestes funcions ja s'eliminen les dades buides.

Data/Hora	Energia activa total	Setmana de l'any	Dia de l'any	Dia de la setmana	Hora del dia	Periode	Setmanes fins examens	Tipus horari biblio
20/3/18 19:00	0.00	12	79	2	19.0	1	2	1
20/3/18 19:15	0.03	12	79	2	19.25	1	2	1
20/3/18 19:30	0.03	12	79	2	19.5	1	2	1
20/3/18 19:45	0.01	12	79	2	19.75	1	2	1
20/3/18 20:00	0.01	12	79	2	20.0	1	2	1
20/3/18 20:15	0.0	12	79	2	20.25	1	2	1
20/3/18 20:30	0.0	12	79	2	20.5	1	2	1
21/3/18 8:45	0.01	12	80	3	8.75	1	2	1
21/3/18 9:00	0.02	12	80	3	9.0	1	2	1
21/3/18 9:15	0.01	12	80	3	9.25	1	2	1
21/3/18 9:30	0.02	12	80	3	9.5	1	2	1
21/3/18 9:45	0.03	12	80	3	9.75	1	2	1

Figura 19. Primeres files del fitxer final de la preparació de les dades

8.2. Entendre el comportament de les dades

En el treball de fi de grau anomenat “Avaluació del funcionament d’un sistema fotovoltaic aïllat a l’ETSEIB” es va realitzar un anàlisi de la demanda dels endolls de la biblioteca a partir del corrent, la potència i l’energia [29]. Es va veure que tots tres valors arriben als seus màxims en les èpoques d’exàmens i als seus mínims a l’estiu, quan la biblioteca és tancada. Per altra banda, també es va notar com a mesura que s’apropaven els exàmens parcials, la demanda a la biblioteca anava augmentant setmana rere setmana. Finalment, es va observar el patró de comportament que presenta la demanda al llarg d’un dia.

Per tal de poder mostrar d’una forma més visual el comportament de la demanda i la relació que hi ha entre la demanda i les variables escollides, s’ha decidit utilitzar la llibreria matplotlib de Python [20]. Cal tenir en compte que aquests gràfics només representen les dades obtingudes quan la biblioteca està oberta, és a dir, les que es tindran en compte per fer la predicció. Sumen un total de 12998 mostres no buides.

En el gràfic de la Figura 20 es veuen quantes mostres hi ha de cada valor observat de l’energia activa. Es veu clarament que el valor més observat és el de 0 KWh i que com més alt és el valor, menys cops s’observa.

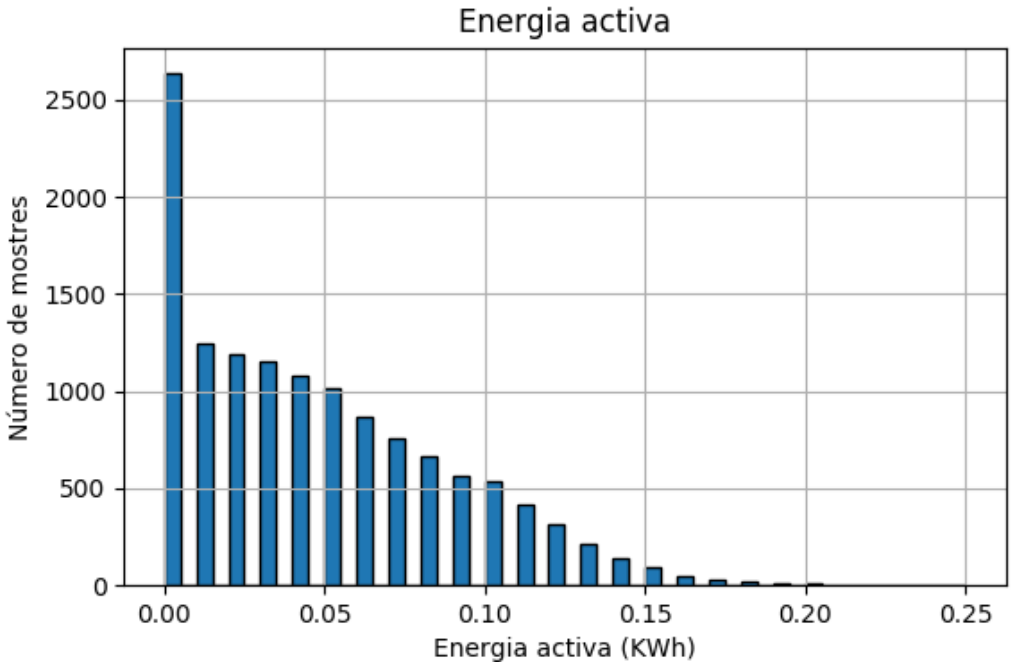

Figura 20. Histograma de les mostres d'Energia activa disponibles

En el gràfic de la Figura 21 es torna a veure la distribució que tenen les mostres d'energia activa, aquest cop en un diagrama de caixa. En aquest es veu clarament el valor màxim observat de l'energia activa: 0,25 KWh. En el capítol 8.3.1. del present treball se'n parla més sobre això.

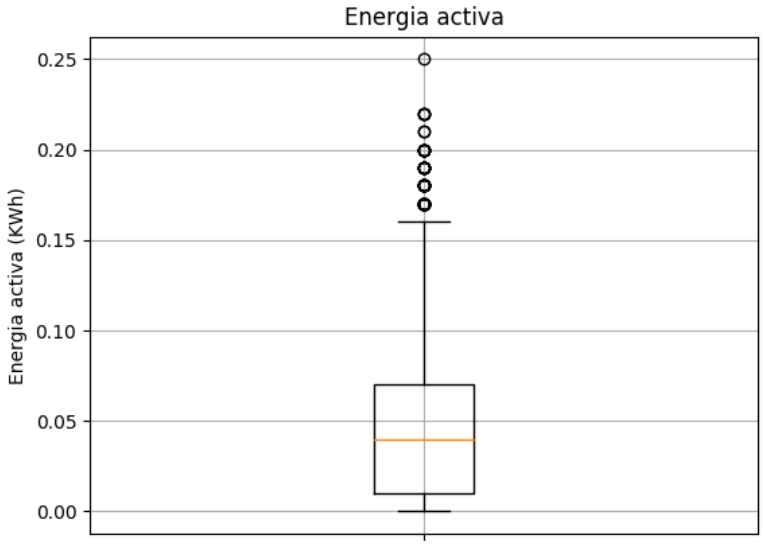

Figura 21. Diagrama de caixa de l'energia activa (KWh)

A la Figura 22 s'observa la relació que hi ha entre el dia de la setmana i la demanda a la biblioteca. Cal tenir en compte que les mostres utilitzades per fer el boxplot de cada dia de la setmana són les del dia en qüestió en el cas que la biblioteca estigués oberta. És per aquest motiu que no és d'estranyar que les mitjanes i els valors més alts siguin els de dissabte i diumenge, ja que la biblioteca només obre els caps de setmana quan és època d'exàmens. En canvi, els boxplots de dilluns a divendres tenen en compte les dades de quasi totes les setmanes de l'any. S'observa que les distribucions de dilluns a dijous no varien massa entre elles, però, en canvi, la mitjana i els valors observats els divendres són molt més baixos.

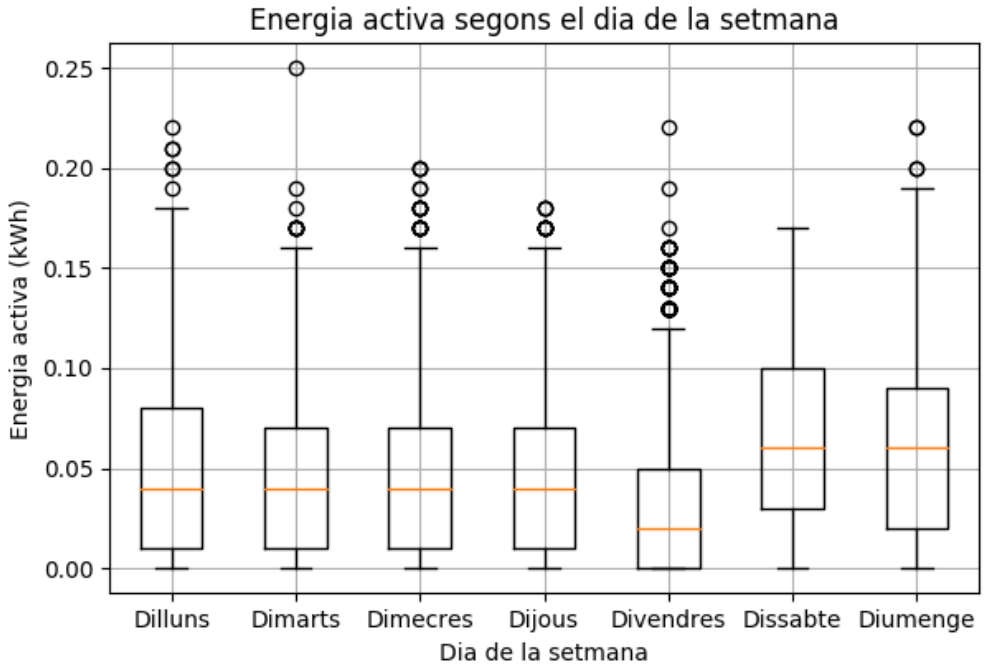

Figura 22. Diagrama de caixes de l'energia activa (KWh) segons el dia de la setmana

A la Figura 23 es vol mostrar la relació entre la demanda observada i el període acadèmic en el que es troba l'escola. Com era d'esperar, els valors més elevats s'observen en les èpoques d'exàmens, especialment durant els finals. També es veu com la demanda un cop passats els parcials també augmenta. L'energia activa durant l'època de reavaluacions és menor a l'observada durant el curs, cosa que s'entén pensant que el número d'estudiants que s'ha de presentar a aquests exàmens també és inferior als matriculats cada quadrimestre. Per últim, en el període anomenat 'Altres' la demanda és quasi nul·la, ja que es tracta de les setmanes festives a nivell acadèmic.

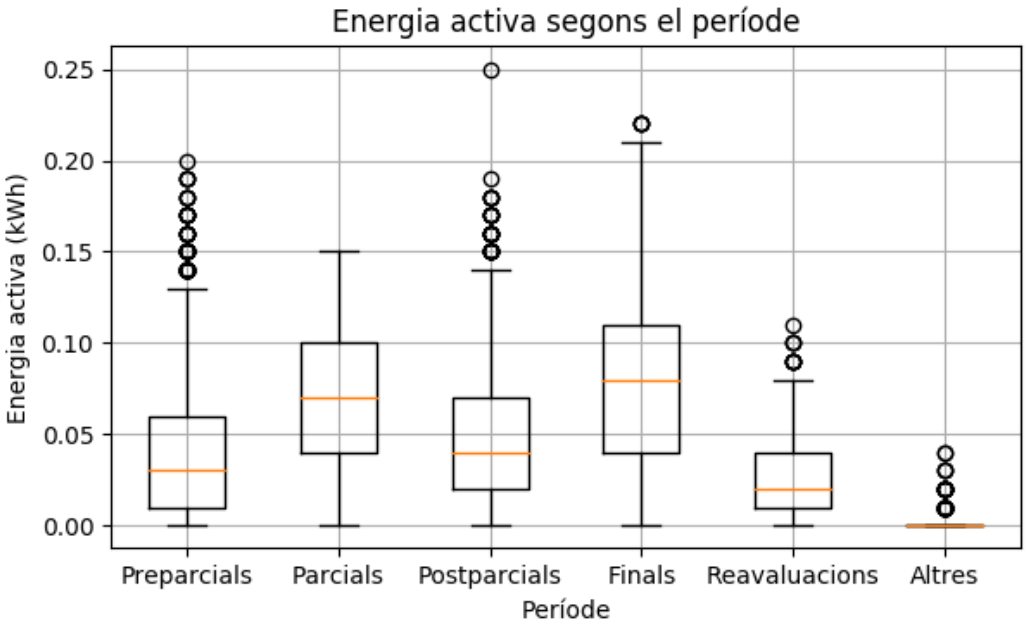

Figura 23. Diagrama de caixes de l'energia activa (KWh) segons el període del curs

En el gràfic de la Figura 24 es pot veure l'energia activa mitjana de cada dia de l'any, en el que es diferencien les èpoques de parcials i finals, així com també les vacances d'estiu i febrer. També és interessant fixar-se que quan no hi ha exàmens i la biblioteca tanca els caps de setmana, la gràfica presenta un perfil de serra. En canvi, en època d'exàmens finals, que la biblioteca obra aproximadament durant un mes sencer, l'energia activa mitjana diària no arriba a ser mai nul·la. Seguint aquest raonament, es pot notar com pels parcials del quadrimestre de primavera la biblioteca obre durant un període més llarg que en els de tardor. Això depèn de si els exàmens parcials de primavera cauen abans o després de setmana santa. Tal i com s'ha comentat en l'apartat 7.2 del present treball, aquest és un factor que no es podrà tenir en compte degut a les dimensions de la base de dades de la que es disposa.

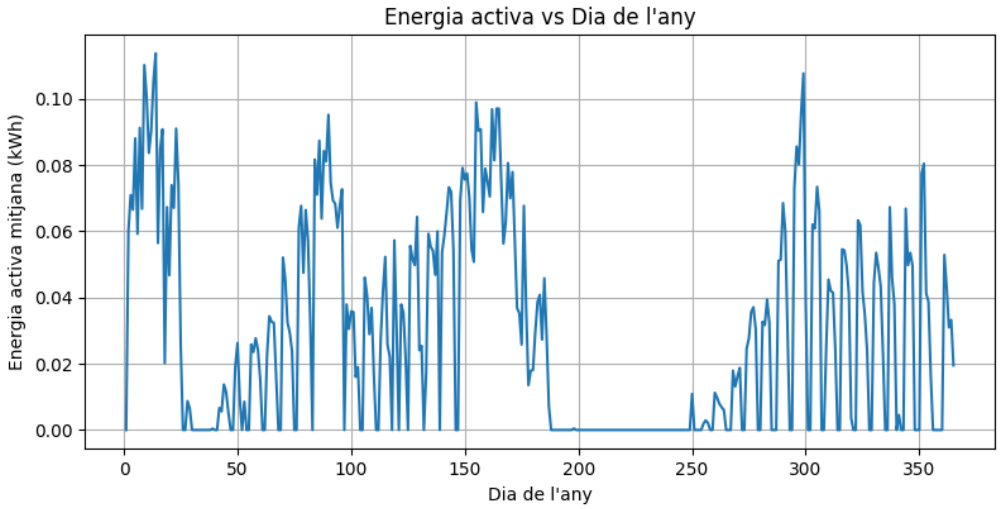

Figura 24. Gràfic de l'energia activa en funció del dia de l'any

En el gràfic de la Figura 25 es pot observar l'energia activa mitjana de cada setmana de l'any. Igual que el gràfic anterior, l'energia activa presenta pics en les èpoques d'exàmens i valors nuls durant les vacances. També es pot veure que a mesura que s'apropen els exàmens, la demanda puja.

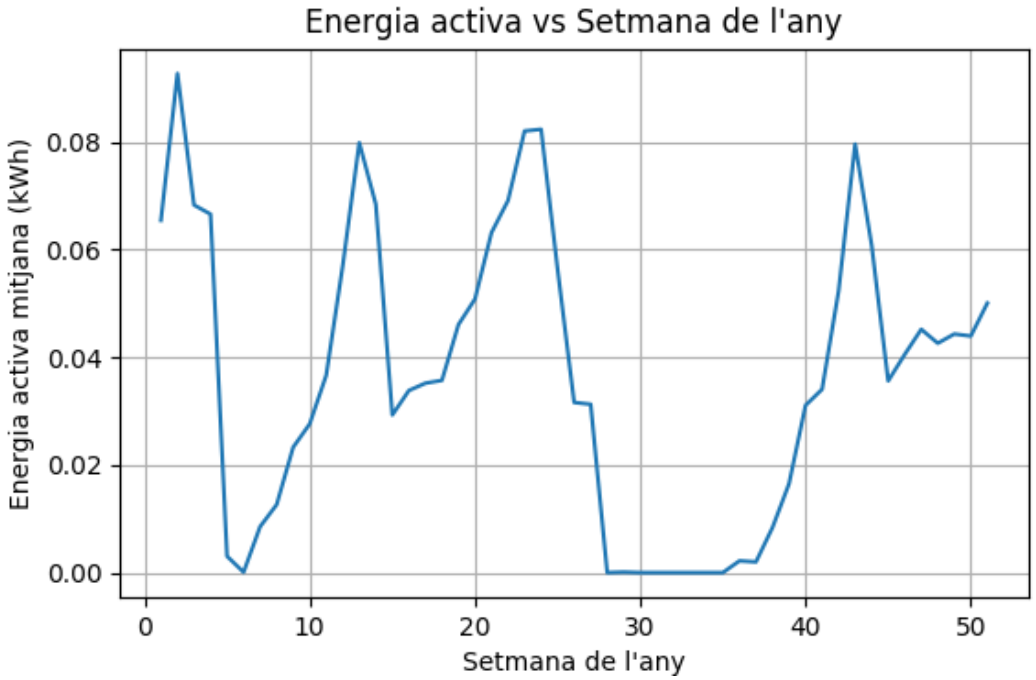

Figura 25. Gràfic de l'energia activa en funció de la setmana de l'any

Finalment, es vol veure el comportament de la demanda segons l'hora del dia, el qual va molt relacionat amb l'horari en el que obre la biblioteca aquell dia. Tot i així, els trets característics són molt semblants. Es veu com a primera hora del dia la demanda no és molt alta i va augmentant a mesura que van arribant els estudiants a la biblioteca o tenen necessitat de carregar algun dispositiu. De la mateixa manera, quan s'apropa l'hora de tancar, la demanda va baixant. També es nota clarament el descans per anar a dinar. A la Figura 26 es mostra l'exemple de quan la biblioteca obre de 8 a 20.30h.

Figura 26. Gràfic de l'energia activa (KWh) en funció de l'hora

Per últim, cal comentar què passa quan la instal·lació es desconnecta perquè les bateries no tenen suficient energia per alimentar els endolls de la sala d'estudis. En aquests casos l'energia activa no queda registrada i per tant no apareixen les corresponents mostres a la base de dades. D'aquesta manera, no representen cap problema a l'hora de crear el model de predicció. Dins del període de la base de dades de la que s'ha disposat, això només va succeir un cop i va ser durant els dies 28 i 29 d'octubre del 2018.

8.3. Preprocessament de les dades

8.3.1. Neteja de les dades

Per una banda, cal mirar si hi ha alguna dada buida o incompleta per tal d'eliminar-la. S'han trobat 3 casos en els que una de les tres línies no va registrar la dada de l'energia activa. Per tal de no alterar la predicció, el millor és eliminar totes les dades corresponents a aquella

hora. En no tractar-se d'un error que es repeteixi seguidament, no cal donar-li més importància.

Per altra banda, en filtrar totes aquelles mostres que presenten una energia activa total major a 0,19 KWh, s'observen els següents valors mostrats a la Figura 27.

```
/Users/anais/PycharmProjects/tfg/venv/bin/python /Users/anais/PycharmProjects/tfg/programa/outlayer.py
            Data/Hora  Energia activa total  ...  Periode  Setmanes fins examens
2580     4/6/18 17:15                  0.20  ...        4                     0
2657     6/6/18 12:30                  0.20  ...        4                     0
2658     6/6/18 12:45                  0.20  ...        4                     0
7690   18/12/18 17:15                  0.25  ...        3                     3
8541    11/1/19 13:00                  0.22  ...        4                     0
8631    13/1/19 11:00                  0.20  ...        4                     0
8632    13/1/19 11:15                  0.22  ...        4                     0
8633    13/1/19 11:30                  0.20  ...        4                     0
8640    13/1/19 13:30                  0.22  ...        4                     0
8704    14/1/19 16:00                  0.20  ...        4                     0
8706    14/1/19 16:30                  0.21  ...        4                     0
8707    14/1/19 16:45                  0.22  ...        4                     0
8708    14/1/19 17:00                  0.21  ...        4                     0
11314   27/3/19 15:45                  0.20  ...        1                     1
```

Figura 27. Llistat de les mostres d'energia activa total major a 0,19 KWh

Es veu com la majoria de dades amb un valor d'energia activa major a 0,19 KWh són registrades en setmanes d'exàmens (siguin parcials o finals), cosa que no sorprèn. També es considera raonable que hi hagi una mostra que presenti 0,20 KWh la setmana anterior als exàmens parcials.

La mostra que cal destacar és la nº 7690, que presenta un valor molt superior als registrats en cap altre moment (0,25KWh) i, a més a més, a tres setmanes de començar exàmens finals. En filtrar les dades del dia en que es va registrar aquesta mostra, es veu com l'energia activa consumida en la mitja hora abans a aquest instant no va ser registrada. A més d'això, el valor de l'energia consumida just després és molt inferior.

La Figura 28 és la llista d'algunes mostres just abans i després del possible error. També es presenta un gràfic de les mostres registrades aquell dia a la Figura 29.

```
7686  18/12/18 15:45                  0.12  ...        3                     3
7687  18/12/18 16:00                  0.12  ...        3                     3
7688  18/12/18 16:15                  0.12  ...        3                     3
7689  18/12/18 16:30                  0.16  ...        3                     3
7690  18/12/18 17:15                  0.25  ...        3                     3
7691  18/12/18 17:30                  0.10  ...        3                     3
7692  18/12/18 17:45                  0.07  ...        3                     3
7693  18/12/18 18:00                  0.10  ...        3                     3
7694  18/12/18 18:15                  0.11  ...        3                     3
7695  18/12/18 18:30                  0.12  ...        3                     3
```

Figura 28. Algunes de les mostres abans i després del possible error trobat

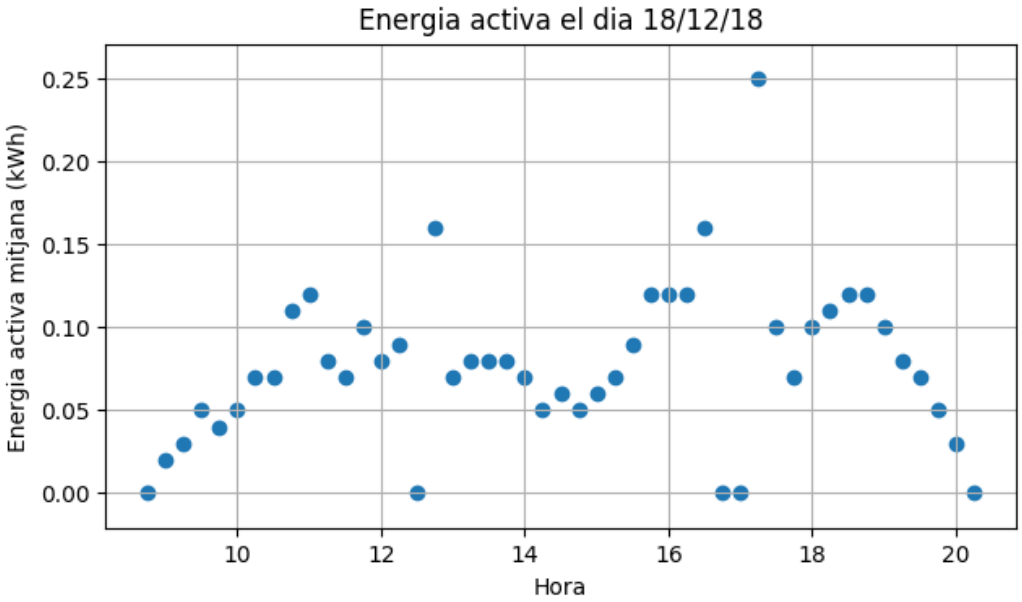

Figura 29. Gràfic de l'energia activa el dia 18/12/18 segons l'hora

Per tots aquests motius, es considera que aquesta mostra presenta un valor excepcional que no representa la demanda de la biblioteca de l'ETSEIB. Per tant, es decideix eliminar aquesta dada i no tenir-la en compte a l'hora de fer la predicció. D'aquesta manera queden un total de 12997 mostres.

8.3.2. Dades d'entrenament i de test

Es decideix separar les dades en un 67% per entrenar i un 33% per testejar.

Per tal de poder calcular de la manera més fiable la qualitat del model final del present projecte, no es faran servir les dades de test fins a l'últim apartat, un cop ja s'ha triat el model definitiu amb els seus paràmetres optimitzats. Per tant, totes els resultats intermedis s'obtindran fent ús únicament de les dades d'entrenament. D'aquesta manera s'evita que les decisions intermèdies es facin basant-se en el comportament que el model té amb les dades de test.

8.3.3. Preparació de les dades

Tal i com s'ha explicat a l'apartat 6. "Metodologia per fer una predicció", la preparació de les dades pot ser útil o no, depenent del comportament que presentin segons el model utilitzat. Les dues tècniques més utilitzades són l'estandardització i la normalització.

Estandarditzar les dades només podria valorar-se com una opció en el cas que la variable que s'està intentant predir (y) o els atributs d'aquesta (X) presentessin una distribució

normal. Tal i com es pot comprovar amb els gràfics de l'Annex B, aquest no és el cas del present projecte. Per tant, es descarta l'opció d'estandarditzar les dades.

En quant a la normalització, tot i que es poden trobar recomanacions sobre si normalitzar les dades segons el model que s'utilitzi, com que el temps d'espera i el cost computacional no són alts, es decideix calcular els resultats normalitzant i sense normalitzar per tal de veure quin mètode ofereix més bons resultats per cada model.

8.4. Elecció i optimització del model sense la variable "Setmanes fins exàmens"

Tal i com s'ha explicat anteriorment, la variable "Setmanes fins exàmens" no té sentit en època de vacances acadèmiques. És per aquest motiu que es decideix crear un primer model de predicció sense tenir en compte aquesta variable.

Més endavant, es crearà un model per aquells dies en que sí que es pugui determinar el valor d'aquesta variable. Aquest segon model tindrà un atribut més, però menys mostres per entrenar i testejar. Caldrà veure si la qualitat de la predicció és més bona en aquest cas.

8.4.1. Comparació de tots els models aplicables

En aquesta part del procés es proven tots aquells algoritmes que ofereix Scikit-learn que s'adeqüen amb la predicció del treball. S'ha fet la tria basant-se amb les opcions presentades a la web oficial de Scikit-learn [34]. Cal tenir en compte que es tracta d'una predicció de regressió i que el conjunt de dades del que es disposa inclou el resultat de la predicció. Dit d'una altra manera, és necessari utilitzar un model de regressió i d'aprenentatge supervisat.

El valor que es decideix utilitzar per fer la comparació dels resultats entre models és el coeficient de determinació R^2. Encara que utilitzant l'error absolut mig i l'error quadràtic mig arribaríem a les mateixes conclusions (sinó molt semblants), el valor R^2 permet entendre més ràpid i fàcilment la qualitat de la predicció. Es tracta d'un coeficient que pot prendre valors des de menys infinit a 1. En els casos de R^2 negatiu, es considera que la mitjana de les variables estima millor el resultat que les funcions establertes per fer la predicció. En canvi, si R^2 dóna un valor entre 0 i 1, aquest és el tant per u de la variabilitat de la variable a predir que s'ha aconseguit explicar gràcies a la predicció feta. [31]

Per tal de fer una primera avaluació dels 30 models considerats, es decideix fer una sola predicció que servirà per saber més o menys quina qualitat es podria arribar a assolir amb cada model. La funció emprada es diu taulacomparacio.py i el seu codi es pot trobar a l'Annex A.

Els resultats obtinguts són els presentats a la Taula 4.

Model	**R^2 Sense preprocés**	**R^2 Normalitzant**	**Millor opció**	**Màxim R^2**
LinearRegression	0,0415	0,0523	Normalització	0,0523
Ridge	0,0415	0,0497	Normalització	0,0497
Lasso	-0,0004	-0,0004	Sense preprocés	-0,0004
ElasticNet	-0,0004	-0,0004	Sense preprocés	-0,0004
Lars	0,0415	0,0523	Normalització	0,0523
LassoLars	-0,0004	-0,0004	Sense preprocés	-0,0004
OrthogonalMatchingPursuit	0,0244	0,0479	Normalització	0,0479
BayesianRidge	0,0415	0,0523	Normalització	0,0523
ARDRegression	-	-	-	-
SGDRegressor	-2,01E+29	0,0229	Normalització	0,0229
PassiveAggressiveRegressor	-0,7374	-1,1899	Sense preprocés	-0,7374
TheilSenRegressor	-0,3197	-2,7351	Sense preprocés	-0,3197
HuberRegressor	0,0242	0,0323	Normalització	0,0323
RANSACRegressor	-1,1265	-900,9544	Sense preprocés	-1,1265
KernelRidge	-0,0234	0,0412	Normalització	0,0412
SVR	-1,9602	-1,9396	Normalització	-1,9396
NuSVR	0,7939	0,1046	Sense preprocés	0,7939
LinearSVR	-0,777	-0,0413	Normalització	-0,0413
KNeighborsRegressor	0,7392	0,5424	Sense preprocés	0,7392
RadiusNeighborsRegressor	-	0,3654	Normalització	0,3654
GaussianProcessRegressor	-20105,0839	0,3654	Normalització	0,3654
PLSRegression	0,0377	0,0453	Normalització	0,0453
PLSCanonical	-0,8349	-3,6211	Sense preprocés	-0,8349
CCA	0,028	-0,036	Sense preprocés	0,028
DecisionTreeRegressor	0,6676	0,6323	Sense preprocés	0,6676
BaggingRegressor	0,7862	0,7694	Sense preprocés	0,7862
RandomForestRegressor	0,7872	0,7716	Sense preprocés	0,7872
ExtraTreesRegressor	0,8273	0,8137	Sense preprocés	0,8273

Model	R^2 Sense preprocés	R^2 Normalitzant	Millor opció	Màxim R^2
AdaBoostRegressor	0,2324	0,1647	Sense preprocés	0,2324
GradientBoostingRegressor	0,6308	0,492	Sense preprocés	0,6308
MLPRegressor	0,0082	0,1266	Normalització	0,1266

Taula 4. Resultats del R^2 (sense preprocés i normalitzant) de la primera comparació entre algoritmes

Es pot veure com no s'han obtingut resultats pel model ARDRegression. Encara que aquest model és aplicable en el cas del present projecte, el temps que triga a entrenar les dades i fer la predicció no és admissible. Es va arribar a esperar un temps de dues hores sense obtenir resultats. És per això que aquest model queda descartat directament.

Per altra banda, també es veu com pel model RadiusNeighborsRegressor no s'ha obtingut cap resultat sense normalitzar les dades. Pel tipus de procés que aquest algoritme aplica, no li és possible realitzar la predicció sense tractar abans les dades.

8.4.2. Comparació dels models finalistes

D'entre tots els models testejats es decideix acceptar com a finalistes els que han obtingut els 5 millors resultats, els quals compleixen $R^2 > 0,7$. Els models seleccionats són els següents:

- NuSVR
- KNeighborsRegressor
- BaggingRegressor
- RandomForestRegressor
- ExtraTreesRegressor

Es veu com en tots 5 casos, encara que sigui per poca diferència, no és necessari fer preprocés. Cap d'ells ofereix un millor comportament si es normalitzen les dades.

Per tal de comparar aquestes 5 opcions es decideix utilitzar el mètode de validació creuada explicat a l'apartat 6. "Metodologia per fer una predicció". D'aquesta manera, en realitzar més d'un test de predicció, no només s'obté un resultat. Així es poden calcular la mitjana i variància dels diferents resultats aconseguits i també del temps necessari per realitzar la predicció. Donada la dimensió de la base de dades, es creu que el millor és dividir les dades en 5 i així obtenir 5 resultats.

La funció emprada s'anomena finalistes.py i el seu codi es troba a l'Annex A. Els resultats obtinguts es presenten a continuació:

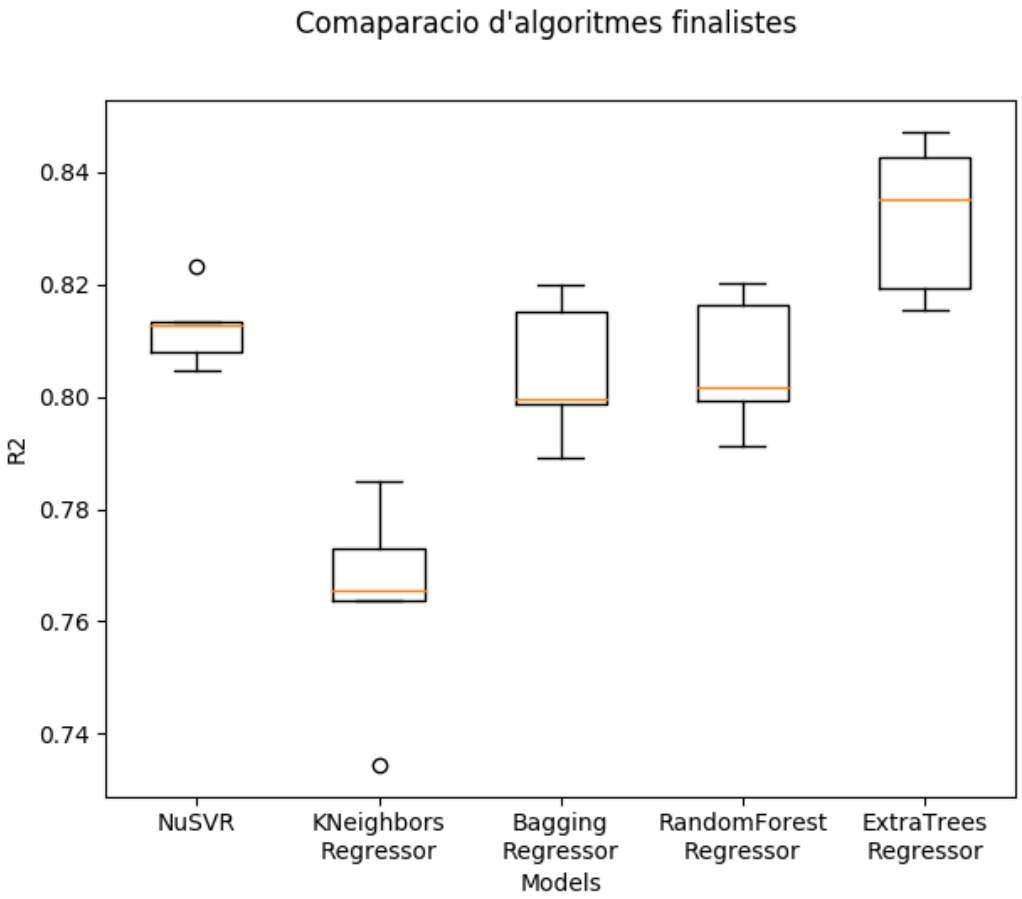

Figura 30. Diagrama de caixes dels resultats del R^2 dels models finalistes

Models	**Mitjana R^2**	**Variància R^2**	**Mitjana temps de predicció (s)**	**Variància temps de predicció**
NuSVR	0,8124	0,0062	0,7372	0,0090
KNeighborsRegressor	0,7644	0,0167	0,0150	0,0003
BaggingRegressor	0,8045	0,0114	0,0111	0,0004
RandomForestRegressor	0,8058	0,0109	0,0107	0,0011
ExtraTreesRegressor	0,8321	0,0125	0,0121	0,0005

Taula 5. Resultats del R^2 i temps de predicció dels models finalistes

Gràcies al gràfic es veu clarament que la millor opció és el model ExtraTreesRegressor, el qual presenta una mitjana de R^2 de 0,8321 i una variància de 0,0125. Encara que la variància de NuSVR és molt inferior, els millors resultats calculats amb aquest model només arriben a tenir la qualitat dels pitjors casos del de ExtraTreesRegressor.

Per altra banda, en quant als temps que triguen a fer una predicció, encara que tots són molt

baixos i no cal descartar cap de les opcions per aquest motiu, és interessant destacar que la segona millor opció per fer la regressió triga 70 vegades el que triga el model ExtraTreesRegressor. Depenent de les aplicacions que se li volgués donar a aquesta predicció, aquestes dècimes de segon podrien marcar la diferència.

8.4.3. Optimització del millor model

Finalment, cal estudiar quin valor han de prendre els paràmetres del model per tal d'obtenir una bona predicció amb un temps de computació acceptable.

Els paràmetres i els corresponents valors predeterminats que presenta el model ExtraTreesRegressor són: [33]

- n_estimators = 'warn'
- criterion = 'mse'
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1
- min_weight_fraction_leaf = 0,0
- max_features = 'auto'
- max_leaf_nodes = None
- min_impurity_decrease = 0,0
- min_impurity_split = None
- bootstrap = False
- oob_score = False
- n_jobs = None
- random_state = None
- verbose = 0
- warm_start = False

Entenen la definició d'aquests, es creu que els valors predeterminats de tots els paràmetres que comencen amb "max_" o "min_" són la millor opció, ja que canviar-los voldria dir limitar el model. Canviar el valor del paràmetre "bootstrap" també implicaria reduir la capacitat de predicció del model, ja que amb el valor actual s'utilitza tota la base de dades per crear cada arbre i no només una part d'aquesta. Com que "oob_score" només pot valdre True quan bootstrap=True, també es deixa amb el valor predeterminat. El paràmetre "n_jobs" indica el número de càlculs que el model realitza en paral·lel. Donat que el temps de computació no és gens elevat, es decideix deixar-lo també amb el valor predeterminat i no realitzar més d'un càlcul al mateix temps. Quant a "random_state", tal i com s'ha explicat anteriorment, s'ha decidit que prengui el valor 7 per totes les funcions del treball. Canviar el paràmetre

"verbosity" faria que quan s'executés la funció es guardessin els passos intermitjos per tal de mostrar-ne els resultats al final, cosa que no és necessari en el present estudi. L'opció "warm_start" permet al model utilitzar els resultats de prediccions que ha fet prèviament per entrenar-se. Encara que aquesta opció pot ser interessant un cop el model està creat, per tal de no influir en l'avaluació final del model i els resultats del treball, es decideix no activar-la i deixar el valor predeterminat.

Per tant, els únics paràmetres que queden per determinat són "n_estimators" i "criterion". "n_estimators" es defineix com el número d'arbres que té el bosc que s'està creant i "criterion" és la funció que mesura la qualitat de la divisió. Per tal de veure els valors òptims d'aquests paràmetres s'utilitza la funció GridSearchCV() que ofereix scikit-learn [32]. Es demana que provi amb els valors n_estimators=[5,10,20,30,50] i les dues opcions disponibles de criterion=['mse','mae'].

La funció emprada s'anomena optimitzacio.py i el seu codi es pot trobar a l'Annex A. S'obtenen els següents resultats:

Posició	**"criterion"**	**"n_estimators"**	**Mitjana R^2**	**Variància R^2**	**Mitjana temps de predicció**	**Variància temps de predicció**
1	mse	50	0,8471	0,0094	0,0407	0,0020
2	mae	50	0,8471	0,0097	0,0430	0,0030
3	mse	30	0,8451	0,0104	0,0253	0,0020
4	mae	30	0,8439	0,0081	0,0250	0,0006
5	mse	20	0,8411	0,0112	0,0180	0,0010
6	mae	20	0,8401	0,0081	0,0183	0,0012
7	mse	10	0,8321	0,0125	0,0104	0,0007
8	mae	10	0,8320	0,0102	0,0103	0,0006
9	mae	5	0,8128	0,0115	0,0065	0,0003
10	mse	5	0,8095	0,0107	0,0062	0,0002

Taula 6. Resultats per la optimització dels paràmetres "criterion" i "n_estimator" del model ExtraTreesRegressor

Analitzant aquests resultats s'arriba a la conclusió que el criteri per valorar la qualitat de la divisió no influencia gaire. Tot i així, com que en la majoria d'opcions de "n_estimators" ha donat un més bon resultat el criteri "mse", es decideix deixar aquest valor pel paràmetre "criterion", que, de fet, és el valor predeterminat.

Per altra banda, s'observa clarament que com més estimadors tingui el model, més bona serà la qualitat de la predicció. Tot i així, també cal tenir en compte el cost computacional que implica un número molt elevat d'estimadors, ja que, tal i com es pot comprovar amb els resultats, a major número d'estimadors, més triga el model a fer la predicció.

Veient que la millora de la qualitat quan es passa de 20 estimadors a 30 és menor al 1%, es decideix optar per fixar el valor de "n_estimators"=20.

En calcular la influència que té cada atribut pel resultat de la predicció amb aquests paràmetres escollits, s'obtenen els resultats de la Taula 7.

Atribut	**Pes de influència (%)**
Hora	38,93
Dia de l'any	17,20
Període del curs	16,94
Setmana de l'any	12,65
Dia de la setmana	8,75
Horari de la biblioteca aquell dia	5,53

Taula 7. Pesos que cada variable té en el model final

Es conclou que la variable que més influencia la demanda de la biblioteca és l'hora del dia (38,93%), que està força lluny de la segona i la tercera variable de més influència: el dia de l'any (17,20%) i el període del curs (16,94%). Les variables que menys influencien el resultat són el dia de la setmana (8,75%) i l'horari de la biblioteca aquell dia (5,53%).

Es pot veure com no hi ha cap variable que influenciï en menys d'un 5%. Tenint en compte aquests resultats i que només es tracta de 6 variables, es decideix considerar-les totes necessàries i no eliminar-ne cap del model.

8.4.4. Model final

Així doncs, el model escollit finalment per fer la predicció de la demanda elèctrica de la biblioteca de l'ETSEIB qualsevol dia que la biblioteca estigui oberta, tant quan hi ha vacances acadèmiques com quan hi ha classes o exàmens, és el següent:

ExtraTreesRegressor amb els paràmetres predeterminats, menys el paràmetre "n_estimators", que es fixa a 20. Els atributs que expliquen millor la demanda amb aquest model són: l'hora, el dia de la setmana, el dia de l'any, la setmana de l'any, el període del curs i l'horari que fa la biblioteca aquell dia.

La qualitat de la predicció que ofereix aquest model partint de les dades d'entrenament, utilitzant com a referència el valor R^2 i fent servir el mètode de validació creuada, es mostra a la Figura 31 i la Taula 8.

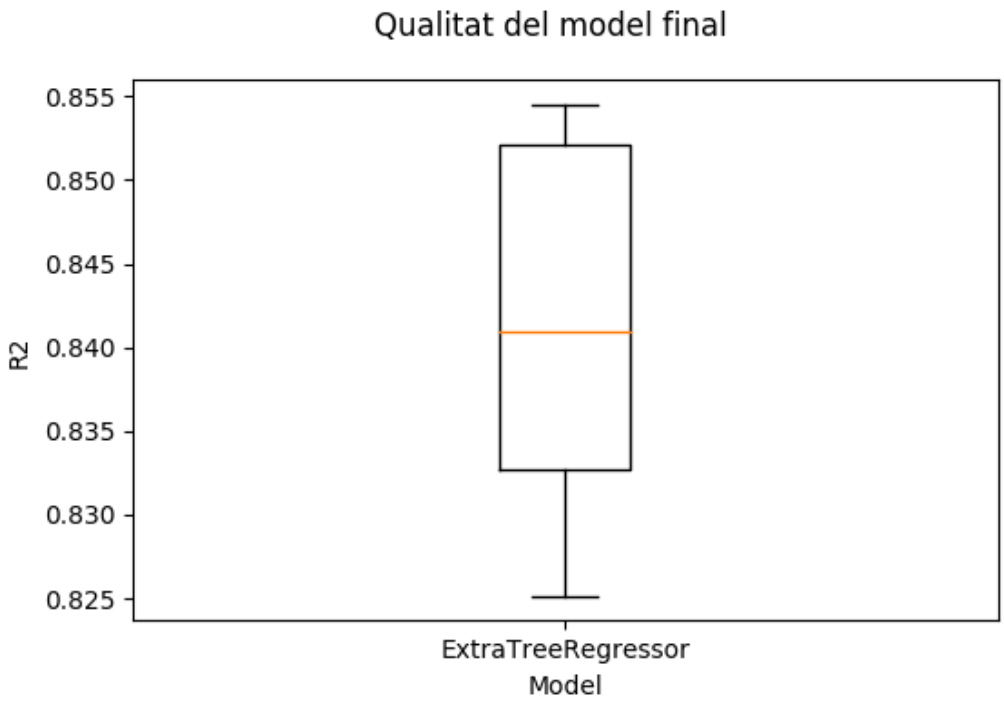

Figura 31. Diagrama de caixa del R^2 del model final

Mitjana R^2	**Variància R^2**	**Mitjana del temps de predicció (s)**	**Variància del temps de predicció**
0,8411	0,0112	0,0191	0,0016

Taula 8. Resultats del R^2 i el temps de predicció del model final

Finalment, un cop triat el model final, cal comprovar quins són els resultats obtinguts de la predicció de les dades de test.

R2 de la predicció del conjunt de dades test	0,8670

Taula 9. Qualitat de la predicció del conjunt de dades de test

8.5. Elecció i optimització del model amb la variable "Setmanes fins exàmens"

Tal i com s'ha comentat, la variable que indica les setmanes que queden fins exàmens només cobra sentit durant el quadrimestre (des de que comencen les classes fins que es fa

l'últim examen final) o durant reavaluacions (des de que es fa l'últim examen final fins que es fa l'últim examen de reavaluació). És a dir, en època de vacances acadèmiques no té sentit parlar de les setmanes que falten fins exàmens.

És per aquest motiu que es decideix crear un model de predicció sense tenir en compte els dies de vacances i introduint la variable "Setmanes fins exàmens". Concretament, els períodes que quedaran fora de l'abast d'aquesta predicció són:

- Les dues setmanes de febrer entre l'últim examen final del quadrimestre de tardor i el primer dia de classe del quadrimestre de primavera.
- Les vacances d'estiu entre l'últim examen de reavaluació i el primer dia de classe del quadrimestre de tardor.

D'aquesta manera, el número de mostres netes es redueix des de 12997 a 12055.

Els passos a seguir seran pràcticament iguals als de l'apartat 8.4. del present treball. A continuació es mostren els resultats.

8.5.1. Comparació de tots els models aplicables

La Taula 10 serveix per comparar entre els 30 models considerats.

Model	**R^2 Sense preprocés**	**R^2 Normalitzant**	**Millor opció**	**Màxim R^2**
LinearRegression	0,1715	0,1623	Sense preprocés	0,1715
Ridge	0,1715	0,1563	Sense preprocés	0,1715
Lasso	-0,0006	-0,0006	Sense preprocés	-0,0006
ElasticNet	-0,0006	-0,0006	Sense preprocés	-0,0006
Lars	0,1715	-1,9022	Sense preprocés	0,1715
LassoLars	-0,0006	-0,0006	Sense preprocés	-0,0006
OrthogonalMatchingPursuit	0,1285	0,1056	Sense preprocés	0,1285
BayesianRidge	0,1715	0,1623	Sense preprocés	0,1715
ARDRegression	-	-	-	-
SGDRegressor	-9,05E+28	0,033	Normalització	0,033
PassiveAggressiveRegressor	-0,7286	-0,536	Normalització	-0,536
TheilSenRegressor	-0,4536	-0,5877	Sense preprocés	-0,4536
HuberRegressor	0,1616	0,1582	Sense preprocés	0,1616
RANSACRegressor	-0,3893	-1,9868	Sense preprocés	-0,3893

Model	R^2 Sense preprocés	R^2 Normalitzant	Millor opció	Màxim R^2
KernelRidge	0,0538	0,1594	Normalització	0,1594
SVR	-1,6678	-1,6239	Normalització	-1,6239
NuSVR	0,7716	0,1622	Sense preprocés	0,7716
LinearSVR	-7,3575	0,1489	Normalització	0,1489
KNeighborsRegressor	0,7132	0,6171	Sense preprocés	0,7132
RadiusNeighborsRegressor	-	0,0031	Normalització	0,0031
GaussianProcessRegressor	-29281,5679	0,4236	Normalització	0,4236
PLSRegression	0,1486	0,1505	Normalització	0,1505
PLSCanonical	-0,5377	-1,9196	Sense preprocés	-0,5377
CCA	0,1316	0,0601	Sense preprocés	0,1316
DecisionTreeRegressor	0,6734	0,6246	Sense preprocés	0,6734
BaggingRegressor	0,7808	0,7579	Sense preprocés	0,7808
RandomForestRegressor	0,7791	0,7625	Sense preprocés	0,7791
ExtraTreesRegressor	0,8152	0,8125	Sense preprocés	0,8152
AdaBoostRegressor	0,3004	0,1341	Sense preprocés	0,3004
GradientBoostingRegressor	0,6069	0,443	Sense preprocés	0,6069
MLPRegressor	-0,1028	0,1325	Normalització	0,1325

Taula 10. Resultats del R^2, sense preprocés i normalitzant, de la primera comparació entre algoritmes

Es tornen a tenir el mateixos problemes pels models ARDRegression i RadiusNeighborsRegressor.

8.5.2. Comparació dels models finalistes

Prenent el mateix criteri que abans, els models que es consideren com a finalistes són els mateixos que en el cas anterior:

- NuSVR
- KNeighborsRegressor
- BaggingRegressor
- RandomForestRegressor
- ExtraTreesRegressor

Es torna a comprovar que el fet de normalitzar les dades no millora el comportament d'aquests models.

Fent una comparació més detallada utilitzant el mètode de validació creuada s'obtenen els resultats mostrats a la Figura 32 i la Taula 11.

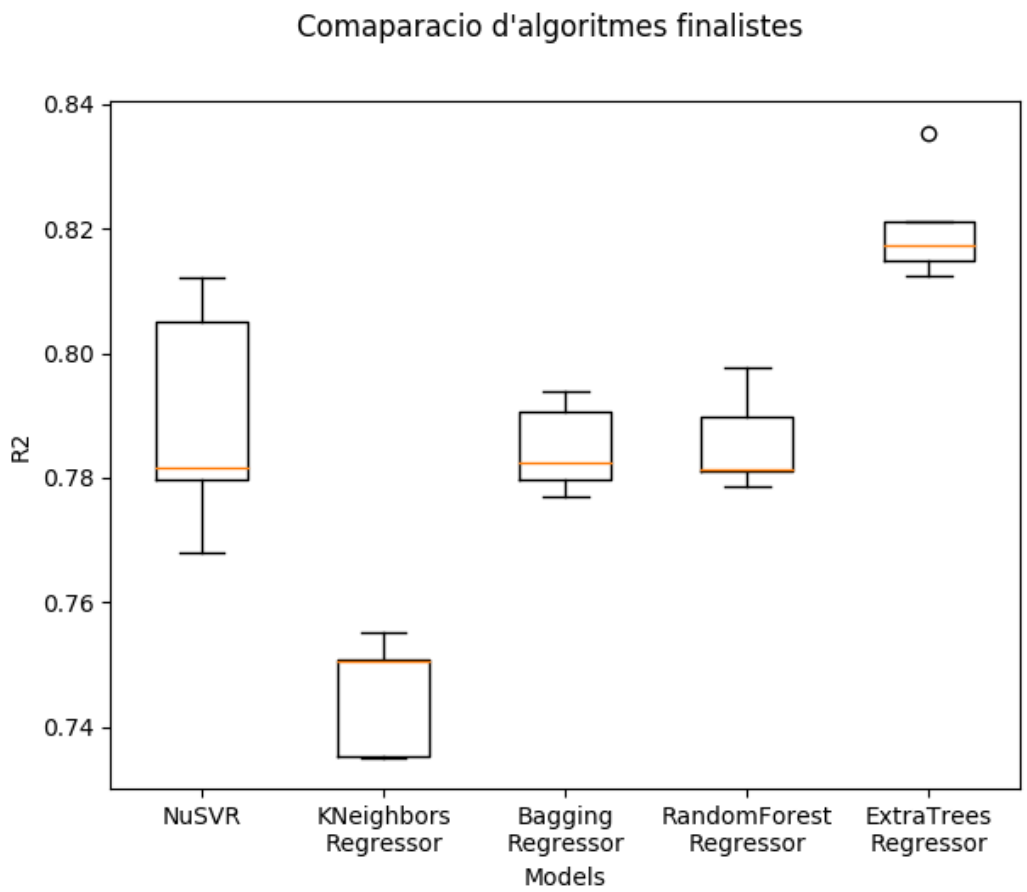

Figura 32. Diagrama de caixes dels resultats del R^2 dels models finalistes

Models	**Mitjana R^2**	**Variància R^2**	**Mitjana temps de predicció (s)**	**Variància temps de predicció**
NuSVR	0,7893	0,0165	0,6282	0,0051
KNeighborsRegressor	0,7454	0,0085	0,0149	0,0002
BaggingRegressor	0,7874	0,0064	0,0103	0,0004
RandomForestRegressor	0,7857	0,0071	0,0098	0,0004
ExtraTreesRegressor	0,8202	0,0082	0,0104	0,0002

Taula 11. Resultats del R^2 i temps de predicció dels models finalistes

De nou, es veu clarament que la millor opció és el model ExtraTreesRegressor, el qual presenta una mitjana de R^2 de 0,8202 i una variància de R^2 de 0,0082.

8.5.3. Optimització del millor model

A partir dels resultats obtinguts en la optimització de la predicció per qualsevol dia que la biblioteca estigui oberta (veure apartat 8.4.3. del present treball), es decideix utilitzar la funció GridSearchCV() per triar el número d'estimadors entre 10, 20 i 30. Els resultats obtinguts són els següents:

Posició	"n_estimators"	Mitjana R^2	Variància R^2	Mitjana temps de predicció (s)	Variància temps de predicció
1	30	0,8350	0,0070	0,0237	0,0018
2	20	0,8311	0,0084	0,0177	0,0012
3	10	0,8202	0,0082	0,0099	0,0006

Taula 12. Resultats per la optimització del paràmetre "n_estimator" del model ExtraTreesRegressor

Aplicant el mateix criteri que abans, es torna a comprovar que la millora de la qualitat entre fixar "n_estimators" a 20 o 30 és inferior a un 1%. Així doncs, es tria optar un altre cop per l'opció de "n_estimators"=20.

Amb aquests paràmetres escollits, es calcula la influència que té cada atribut al resultat de la predicció i s'obtenen els resultats de la Taula 13.

Atribut	Pes de influència (%)
Hora	43,23
Setmanes fins exàmens	17,74
Dia de l'any	12,36
Dia de la setmana	10,22
Setmana de l'any	6,46
Període del curs	6,39
Horari de la biblioteca aquell dia	3,59

Taula 13. Pesos que cada variable té en el model final

Per aquest model també es conclou que la variable que més influencia la demanda de la biblioteca és l'hora del dia (43,23%), inclús amb un tant per cent superior que en el model

anterior. La segona variable que més influencia és el número de setmanes que falten fins exàmens (17,74%). Seguidament, el dia de l'any (12,36%) i el dia de la setmana (10,22%). La variable que menys influencia torna a ser un altre cop l'horari de la biblioteca aquell dia (3,59%).

Encara que la variable que indica l'horari que fa la biblioteca aquell dia és només d'un 3,59%, tenint en compte que no hi ha gaires atributs, es decideix considerar totes les variables necessàries i no eliminar-ne cap del model.

8.5.4. Model final

En el cas d'utilitzar l'atribut que indica les setmanes que falten fins exàmens, el millor model és el següent:

ExtraTreesRegressor amb els paràmetres predeterminats, menys el paràmetre "n_estimators", que es fixa a 20. Els atributs que expliquen millor la demanda amb aquest model són: l'hora, el dia de la setmana, el dia de l'any, la setmana de l'any, el període del curs, l'horari que fa la biblioteca aquell dia i el número de setmanes fins exàmens.

La qualitat de la predicció que es veu que ofereix aquest model partint de les dades d'entrenament, utilitzant com a referència el valor R^2 i fent servir el mètode de validació creuada, es mostra en el gràfic de la Figura 33 i en la Taula 14.

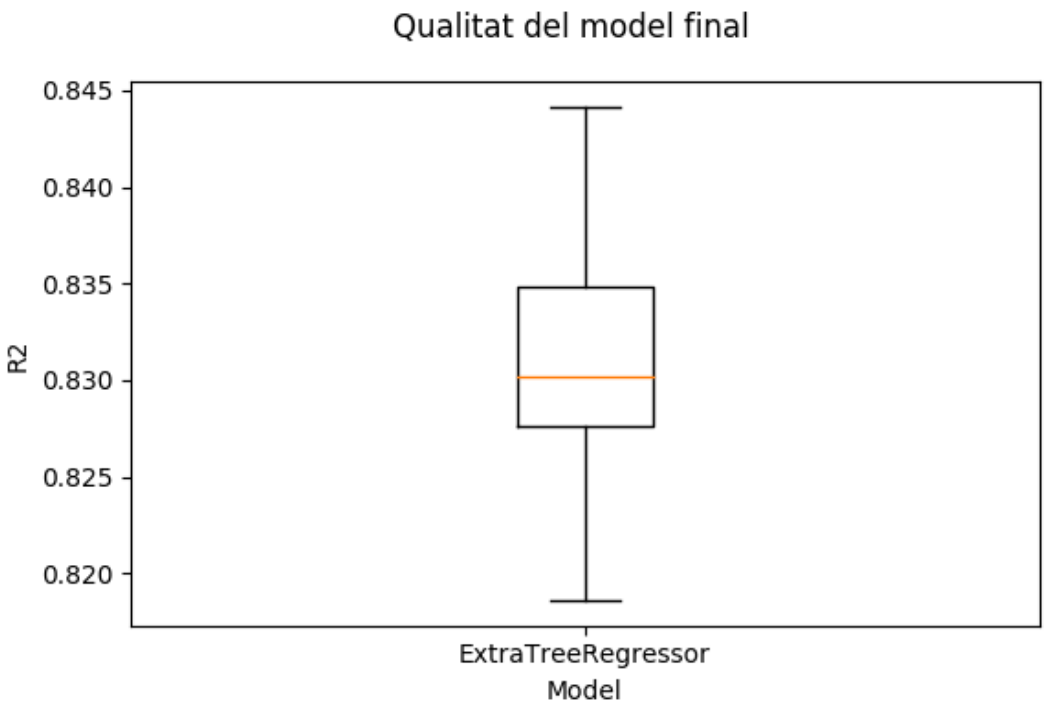

Figura 33. Diagrama de caixa del R^2 del model final

Mitjana R^2	Variància R^2	Mitjana del temps de predicció (s)	Variància del temps de predicció
0,8311	0,0084	0,0209	0,0049

Taula 14. Resultats del R^2 i el temps de predicció del model final

Finalment, un cop triat el model final, cal comprovar quins són els resultats obtinguts de la predicció de les dades de test.

R2 de la predicció del conjunt de dades test	0,8506

Taula 15. Qualitat de la predicció del conjunt de dades de test

8.6. Conclusió dels resultats

8.6.1. Elecció final del model

És interessant veure com el model que inclou la variable "Setmanes fins exàmens", encara que per poca diferència, dóna una mitjana del valor R^2 inferior al que ens ofereix el model que no contempla les setmanes fins exàmens. Tot i així, és important fixar-se que, de totes maneres, aquesta variable és la segona que ofereix més informació i té més pes a l'hora de fer la predicció.

Aquest fet podria tenir vàries explicacions. En primer lloc, cal tenir en compte la reducció en el nombre de mostres amb les que s'ha entrenat el model. Per altra banda, també es pot explicar pensant que la predicció de l'energia activa quan no hi ha exàmens o classes obté molts més bons resultats. Això fa que la mitjana dels resultats en aquest model baixi una mica.

Per tant, encara que el darrer model pot semblar en un primer moment pitjor que l'anterior, es considera que, per tal de predir la demanda elèctrica a la biblioteca de l'ETSEIB en períodes de classes o exàmens, cal tenir en compte l'atribut de les setmanes que queden fins exàmens.

També cal definir el millor model per predir la demanda elèctrica quan no hi ha classes o exàmens, és a dir, durant les dues setmanes de febrer que hi ha festa entre el quadrimestre de tardor i el de primavera o durant les setmanes d'estiu un cop ja s'han fet les reavaluacions. En aquests dos casos no té sentit parlar de la variables "Setmanes fins exàmens" i, per tant, el model no pot comprendre aquest atribut.

8.6.2. Limitacions dels models de predicció presentats

Cal tenir present que els models de predicció proposats tenen unes quantes limitacions, principalment per la reduïda base de dades de la que s'ha disposat, la qual conté informació de poc més d'un any.

Això implica que es perd informació sobre la variabilitat al llarg dels anys. Per exemple, no s'ha pogut tenir en compte la influència que té si els parcials del segon quadrimestre cauen abans o després de setmana santa. Pel mateix motiu, tampoc s'ha pogut valorar si la quantitat d'estudiants matriculats a l'escola aquell curs o quadrimestre influeix.

Per tant, si s'intentessin aplicar els models presentats durant un altre curs, el més probable és que la qualitat de la predicció no arribés als nivells calculats.

També a causa de la limitada base de dades, hi ha valors d'atributs que presenten un número molt reduït de mostres. Això es pot comprovar amb els gràfics de l'Annex B. Per exemple, es pot veure com el número de mostres de 20 a 22h és molt més reduït que els altres. De la mateixa manera, hi ha alguns horaris de biblioteca que no presenten gaires exemples. Això provoca que les conclusions a les que s'arriba en aquests casos no són gaire fiables.

9. Impacte ambiental

Existeix un ampli consens en que la situació mediambiental global és molt greu. És feina de tots, tant països, associacions, empreses i particulars, prendre mesures per tal de posar remei als problemes que s'han anat agreujant al llarg dels anys, especialment en les últimes dècades. Cal tenir-ho present en qualsevol activitat que es realitzi, tant en hores de lleure com quan s'està treballant.

L'efecte directe que té aquest treball al medi ambient és pràcticament nul. Com a molt es podria valorar l'energia elèctrica consumida per l'ordinador durant les hores de treball, ja que no s'ha imprès la memòria en cap moment.

A més a més, no es pot oblidar l'ajuda que suposarien els resultats d'aquest treball per aconseguir una gestió òptima de les bateries en el cas que es seguís apostant pels projectes relacionats amb la instal·lació fotovoltaica de l'ETSEIB.

Concretament, es podria aconseguir alimentar més aparells amb el mateix nombre de panells i mateixa capacitat de les bateries. Per exemple, es podrien crear més llocs de treball a la sala d'estudi de la biblioteca o aprofitar l'energia per alimentar altres aparells. D'aquesta manera s'estaria estalviant consumir energia elèctrica de la xarxa.

10. Pressupost

Donat que la instal·lació ja estava muntada i en funcionament quan es va iniciar el present treball, no s'han tingut en compte les despeses que va suposar posar en marxa la instal·lació, ni tampoc els costos que té mantenir-la.

Per calcular el pressupost del present projecte, s'ha dividit el volum de despeses entre recursos humans i recursos informàtics.

Per les despeses de recursos humans, s'ha dividit la feina feta en tres activitats principals:

- Investigació: Consisteix en informar-se de l'estat de l'art de l'energia fotovoltaica i l'anàlisi predictiu, així com també de les característiques de la instal·lació fotovoltaica de l'ETSEIB. Per altra banda, també s'inclouen aquí les hores dedicades a entendre les diferents funcions i possibilitats que dóna Python, Pandas, Scikit Learn i Matplotlib.
- Predicció: Es podria dividir en les tasques d'obtenció de dades, programació i anàlisi de resultats. Vindria a ser la part pràctica del treball.
- Redacció de la memòria

L'estimació dels costos humans es mostra a la següent taula:

	Hores invertides (h)	**Preu per hora (€/h)**	**Preu total (€)**
Investigació	100	15	1.500,00
Predicció	150	15	2.250,00
Redacció de la memòria	80	15	1.200,00
Subtotal	330		4.950,00
IVA (21%)			1039,50
Total			**5989,50**

Taula 16. Estimació dels costos humans

En quant als recursos informàtics, només s'ha tingut en compte l'ús que s'ha fet de l'ordinador portàtil MacBook, el qual té un processador de 2,4 GHz Intel Core 2 Duo i una memòria de 6 GB 1067 MHz DDR3. Cal tenir en compte que tots els programes utilitzats són de lliure llicència.

L'estimació de costos informàtics es presenta a la següent taula:

	Preu (€)	**Amortització (anys)**	**Temps d'ús (mesos)**	**Cost (€)**
Ordinador portàtil	2000	6	4	111,11
Subotal				111,11
IVA (21%)				23,33
Total				**134,44**

Taula 17. Estimació dels costos informàtics

Per tant, sumant els costos dels recursos humans (5989,50€) i els dels recursos informàtics (134,44€), s'obté el cost total del present projecte: **6123,94€**

Conclusions

Aquest projecte és la continuació d'un conjunt de treballs relacionats amb la instal·lació solar fotovoltaica aïllada de l'ETSEIB. Concretament, és el primer que s'endinsa en el món de l'optimització de la gestió energètica. Amb l'objectiu de que en un futur es pugui optimitzar la gestió energètica de les bateries, intenta trobar el millor model per predir la demanda elèctrica de les sales d'estudi alimentades amb la instal·lació en qüestió.

Per començar s'ha realitzat un estudi de l'estat de l'art de l'energia fotovoltaica i l'anàlisi predictiu, i seguidament s'ha intentat comprendre el funcionament de la instal·lació fotovoltaica aïllada de l'ETSEIB. També s'ha explicat la metodologia general seguida per fer una predicció.

Abans de començar amb l'anàlisi de resultats, s'han descrit els trets característics de la predicció objecte d'estudi: el tipus de predicció, la variable que s'està intentant predir i els atributs que influeixen el resultat de la predicció.

La variable que s'ha considerat que representa millor la demanda elèctrica de les sales d'estudi és la suma de l'energia activa (KWh) consumida a les tres línies que presenta la instal·lació. Es tracta d'una predicció d'aprenentatge supervisat, concretament una regressió.

Les variables que s'han considerat des d'un primer moment susceptibles a influenciar el valor del resultat són: l'hora, el dia de la setmana, el dia de l'any, la setmana de l'any, el període del curs acadèmic, l'horari que fa la biblioteca aquell dia i les setmanes que falten fins exàmens.

S'ha vist que pel període del curs acadèmic en el que no hi ha ni classes ni exàmens, no té sentit parlar de quantes setmanes falten fins exàmens. És per aquest motiu que s'ha decidit buscar dos models de predicció:

- Un primer que tingui en compte totes les dades netes disponibles però que no consideri la variable que explica quantes setmanes queden fins exàmens
- Un segon que sí que tingui en compte la variable de quantes setmanes queden fins exàmens i que, per tant, no consideri les mostres fetes quan no hi ha ni classes ni exàmens.

En tots dos casos s'han aconseguit uns models de predicció amb un coeficient de determinació R^2 entre el 0,82 i el 0,85.

La qualitat de la predicció obtinguda sembla satisfactòria, però cal tenir en compte les

limitacions que presenta el model. Degut a la limitada base de dades, que només disposa de mostres durant poc més d'un curs acadèmic, els models presentats no tenen en compte la variabilitat entre anys i, per tant, si s'apliquessin durant un altre curs, el més probable és que no s'obtinguessin uns resultats fiables.

Possibles tasques futures

Per tal de superar aquestes limitacions, es proposa que, de cara a un futur i utilitzant una base de dades més àmplia, es validi i acabi de completar el model escollit per tal de millorar-lo i obtenir una predicció més exacte i precisa.

Cal tenir en ment que, per tal de millorar la gestió energètica de les bateries, és necessari crear un model de predicció per saber la demanda d'energia dels llocs de treball de la biblioteca i un altre model per predir la generació d'energia elèctrica. Un cop aconseguits aquests dos models, es podria passar a estudiar com gestionar la càrrega de les bateries de la manera més òptima.

Per tant, es proposa també realitzar un treball en el que es creï un model de predicció de la generació d'energia elèctrica depenent de l'època de l'any, la previsió de núvols i altres variables que es puguin considerar d'influència. Aquesta model es podria trobar també amb les eines d'aprenentatge computacional que han sigut utilitzades per aquest treball.

Finalment, com a última proposta, estaria la d'optimitzar la gestió de les bateries tenint en compte la predicció de la demanda energètica a la biblioteca i la generació d'energia dels panells fotovoltaics. L'objectiu seria que, gràcies a aquesta optimització, les bateries de la instal·lació no es quedessin mai sense energia suficient per alimentar la demanda existent. Fins i tot es podria plantejar la creació de més llocs de treball.

Agraïments

En primer lloc i de forma principal, m'agradaria agrair l'ajuda rebuda pel tutor Roberto Villafàfila i la confiança dipositada en mi des d'un primer moment per realitzar aquest treball. Els consells i ànims rebuts han sigut essencials pel desenvolupament del projecte.

També m'agradaria donar les gràcies a la Sara Barja Martínez per introduir-me en el món de les prediccions amb aprenentatge computacional, així com també per l'assessorament rebut.

Per altra banda, agrair a la biblioteca de l'ETSEIB la facilitat per obtenir els calendaris d'anys anteriors.

Finalment, no podia faltar l'agraïment pel suport incondicional que representa la meva família. Als meus pares, per mostrar sempre interès en allò que faig i confiar plenament en mi. També al meu germà, que tot just començant la mateixa carrera, desprèn més energia que ningú.

Gràcies també a tots els amics i amigues amb els que he compartit moltes hores de biblioteca i algunes al bar. Ells són els responsables del meu bon humor i han sigut el motor que m'ha mantingut motivada i amb forces per tot.

Bibliografia

Referències bibliogràfiques

[1] AMT SOLAR. *¿Qué es la energía fotovoltaica?* Consulta: Març 2019. Disponible a: http://www.amt-solar.com/index.php/es/fotovoltaica

[2] AQUALIA. *Compromiso con la innovación.* Consulta: Maig 2019. Disponible a: http://compromisoreal.com/compromiso-con-la-innovacion

[3] ASOCIACIÓN DE EMPRESAS DE ENERGÍAS RENOVABLES (APPA). *¿Qué es la energía fotovoltaica?* Consulta: Març 2019. Disponible a: https://www.appa.es/appa-fotovoltaica/que-es-la-energia-fotovoltaica/

[4] ATERSA. *Ficha técnica módulos fotovoltaicos A-240P / A-245P / A-250P (TYCO 3.2).* Consulta: Abril 2019. Disponible a: http://www.atersa.com/Common/pdf/atersa/manuales-usuario/modulos-fotovoltaicos/Ficha%20Tecnica%20A-240P%20-%20A-250P%20Ultra.pdf

[5] BIBLIOTÈCNICA. *Biblioteca de l'Escola Tècnica Superior d'Enginyeria Industrial de Barcelona.* Consulta: Abril 2019. Disponible a: https://bibliotecnica.upc.edu/etseib

[6] CENIT SOLAR. *Fotovoltaica aislada. Esquema de principio.* Consulta: Març 2019. Disponible a: http://www.cenitsolar.com/fotovoltaica_esquema.php

[7] CIRCUTOR. *CirPower.* Consulta: Abril 2019. Disponible a: http://circutor.es/es/productos/energias-renovables/autoconsumo-con-acumulacion-con-conexion-a-red/cirpower-detail

[8] CIRCUTOR. *EDS-Delux*. Consulta: Abril 2019. Disponible a: http://circutor.es/es/productos/medida-y-control/sistemas-de-control/gestor-energetico/servidor-eds-deluxe-efficiency-data-server-detail

[9] CIRCUTOR. *PowerStudio Scada Deluxe.* Consulta: Abril 2019. Disponible a: http://circutor.es/es/productos/medida-y-control/software-de-gestion-energetica/software-power-studio-scada-deluxe-detail

[10] CIRCUTOR. *Serie CVM-1D.* Consulta: Abril 2019. Disponible a: http://circutor.es/es/productos/medida-y-control/analizadores-de-redes-fijos/analizadores-de-redes/serie-cvm-1d-detail

[11] CIRCUTOR. *Servidor web del PowerStudio Scada aplicat a la instal·lació fotovoltaica aïllada de l'ETSEIB.* Consulta: Març 2019. Disponible des de la xarxa interna de la UPC a: http://10.70.0.124/html5/index.html

[12] CIRCUTOR. *Servidor web del Efficiency Data Server.* Consulta: Maig 2019. Disponible des de la xarxa interna de la UPC a: http://10.70.0.123/

[13] DÍAZ, T.; CARMONA, G. *Instalaciones solares fotovoltaicas.* McGraw-Hill Interamericana de España S.L. (Madrid, Abril 2010)

[14] ESPINO, C. *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo – herramientas Open Source que permiten su uso.* Universitat Oberta de Catalunya (Gener 2017).

[15] ETSEIB. *Calendaris ETSEIB.* Consulta: Abril 2019. Disponible a: https://etseib.upc.edu/ca/estudis/calendaris

[16] FERNÁNEZ JIMÉNEZ, D. *Modelo de predicción de la demanda eléctrica horaria a muy corto plazo: aplicación del sistema peninsular español.* Industriales ETSII UPM. (Madrid, Junio 2016).

[17] FREDDY GODOY VIERA, A. *Técnicas de aprendizaje de máquinas utilizadas para la minería de texto.* Investigación Bibliotecológica, vol3.1, núm. 71 (Mèxic, Gener/Abril 2017)

[18] MATHWORKS. *Análisis predictivo. Tres cosas que es necesario saber.* Consulta: Maig 2019. Disponible a: https://es.mathworks.com/discovery/predictive-analytics.html

[19] MATHWORKS. *Machine learning. Tres cosas que es necesario saber.* Consulta: Maig 2019. Disponible a: https://es.mathworks.com/discovery/machine-learning.html

[20] MATPLOTLIB. Consulta: Maig 2019. Disponible a: https://matplotlib.org

[21] MÉNDEZ, FRAN. *¿Cómo el Big Data ayudó a Obama a ganar?* Forbes. (Septiembre 2015). Consulta: Maig 2019. Disponible a: http://forbes.es/emprendedores/7560/como-el-big-data-ayudo-a-obama-a-ganar/

[22] MORERA, B. *Monitorització d'una instalación fotovoltaica.* UPCommons (Abril 2018).

[23] NCSS Statistical software. *Linear Regression and Correlation.* Consulta: Maig 2019. Disponible a: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Linear_Regression_and_Correlation.pdf

[24] PANDAS. Consulta: Maig 2019. Disponible a: https://pandas.pydata.org

[25] PASTOR, M. *Instal·lació solar fotovoltaica per a l'alimentació d'ordinadors portàtils i telefòns mòbils dels usuaris d'una biblioteca.* UPCommons (Gener 2016).

[26] PYTHON SOFTWARE FOUNDATION. Consulta: Maig 2019. Disponible a: https://www.python.org

[27] PYTHON SOFTWARE FOUNDATION. *datetime – Basic date and time types.* Consulta: Abril 2019. Disponible a: https://docs.python.org/2/library/datetime.html

[28] RED ELÉCTRICA DE ESPAÑA. *Demanda de energía eléctrica en tiempo real.* Consulta: Maig 2019. Disponible a: https://demanda.ree.es/visiona/peninsula/demanda/total

[29] SÁNCHEZ MASSANEDA, P. *Avaluació del funcionament d'un sistema fotovoltaic aïllat a l'ETSEIB.* UPCommons (Gener 2019).

[30] SCIKIT LEARN. Consulta: Maig 2019. Disponible a: https://scikit-learn.org/stable/

[31] SCIKIT LEARN. *sklearn.metrics.r2_score.* Consulta: Abril 2019. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[32] SCIKIT LEANR. *sklearn.model_selection_GridSearchCV*. Consulta: Abril 2019. Disponible a: https://scikit-

learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[33] SCIKIT LEARN. *sklearn.ensemble.ExtraTreesRegressor.* Consulta: Maig 2019. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html

[34] SCIKIT LEARN. Supervised learning. Consulta: Març 2019. Disponible a: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[35] SUNFIELDS. *Tipos de paneles solares en el sector fotovoltaico. Clasificación de los tipos de placas solares por su tecnología.* Consulta: Març 2019. Disponible a: https://www.sfe-solar.com/noticias/articulos/tipos-de-paneles-solares-fotovoltaicos/

[36] TABSPAIN. *TAB Solar TOPzS.* Consulta: Abril 2019. Disponible a: https://www.tabspain.com/renovables/solar/baterias-topzs/

[37] UPC. UNIVERSITAT POLITÈCNICA DE CATALUNYA. *S'inicien els treballs del projecte BISOL a la Biblioteca de l'ETSEIB – UPC Energia 2020 – Comunitats sostenibles.* Consulta: Març 2019. Disponible a: https://www.upc.edu/energia2020/ca/noticies/sinicien-els-treballs-del-projecte-bisol-ala-biblioteca-de-letseib

[38] ZAMORANO RUIZ, J. *Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea.* Universidad complutense de Madrid (Madrid, Julio 2018).

Annex A: Codis de les funcions emprades

A.1. Funció afegirvariables.py

```
import datetime
import biblioooberta
import calendariacademic

finicial='fitxerinicial.csv'
ffinal='atributs.csv'

dfini = open(finicial, 'r')
dffinal = open(ffinal, 'w')

n=0
for e in dfini:
    e = e[:-1]
    if n==0:
        dffinal.write("Data/Hora;Energia activa total;Setmana de l'any;Dia de
l'any;Dia de la setmana;Hora del dia;Biblioteca oberta;Periode;Setmanes fins
examens;Tipus horari biblio\n")
    else:
        l = e.split(';')

        en1=l[1].replace(',', '.')
        en2=l[2].replace(',', '.')
        en3=l[3].replace(',', '.')
        EnActTotal=0
        if en1!='':
            EnActTotal=EnActTotal+float(en1)
        if en2 != '':
            EnActTotal = EnActTotal + float(en2)
        if en3!='':
            EnActTotal=EnActTotal+float(en3)
        EnActTotal=round(EnActTotal,2)

        [data,horamin]=l[0].split(' ')
        [dia,mes,any]=data.split('/')
        [hora,min]=horamin.split(':')

        dia=int(dia)
        mes=int(mes)
        any=int(any)+2000
        hora=int(hora)
        min=int(min)

        data=datetime.date(any,mes,dia)

        setmanadelany=data.isocalendar()[1]

        iniciany=datetime.date(any-1,12,31)
        diadelany = (data - iniciany).days
```

```
        #en cas que sigui un any de traspàs
        desembre31=datetime.date(any,12,31)
        numdiesquetelany=(desembre31-iniciany).days
        if numdiesquetelany==366 and data>datetime.date(any,2,28):
                diadelany=diadelany-1

        diadelasetmana=data.isocalendar()[2]

        horadeldia=hora+min/60

        [biblioteca,tipushorari]= biblioooberta.biblioooberta(data, horadeldia)

        [periode,setmanesfinsexams]= calendariacademic.calendariacademic(data,
setmanadelany)

        dffinal.write(l[0] + ';' + str(EnActTotal) + ';' + str(setmanadelany) +
';' + str(diadelany) + ';' + str(diadelasetmana) + ';' + str(horadeldia) + ';'
+ str(biblioteca) + ';' + str(periode) + ';' + str(setmanesfinsexams) + ';' +
str(tipushorari) + '\n')

    n=n+1


dfini.close()
dffinal.close()
```

A.2. Funció biblioooberta.py

```
import datetime

# Llistes de les setmanes que són de cada tipus. Quan la setmana és +100 és pq
es del 2019
llsetm0=[31,32,33,34,35,52,116]
llsetm1=[10,11,12,14,15,16,17,18,19,20,21,22,23,24,25,26,27,37,38,39,40,41,42,4
3,44,45,46,47,48,49,50,51,102,103,104,105,106,107,108,109,110,111,112,115,117,1
18,119,120,121,122]
llsetm2=[28,29,30,36]
llsetm3=[101]
llsetm4=[113,114]
llsetm7=[13]
llposssetm=[llsetm0,llsetm1,llsetm2,llsetm3,llsetm4,[],[],llsetm7]

# Llista amb els dies que son excepcio
lldia0=[datetime.date(2018,5,1),datetime.date(2018,5,21),datetime.date(2018,9,1
1),datetime.date(2018,9,24),datetime.date(2018,10,12),datetime.date(2018,12,6),
datetime.date(2019,1,1),datetime.date(2019,4,22),datetime.date(2019,5,1)]
lldia1=[datetime.date(2019,4,5)]
lldia2=[datetime.date(2018,9,10),datetime.date(2018,12,7)]
lldia3=[datetime.date(2018,10,27),datetime.date(2018,10,28),datetime.date(2018,
11,1),datetime.date(2018,12,27),datetime.date(2018,12,28),datetime.date(2018,12
,29),datetime.date(2018,12,30),datetime.date(2019,1,5),datetime.date(2019,1,6),
```

```
datetime.date(2019,1,12),datetime.date(2019,1,13),datetime.date(2019,1,19),date
time.date(2019,1,20),datetime.date(2019,3,30),datetime.date(2019,3,31)]
lldia4=[]
lldia5=[datetime.date(2018,7,6)]
lldia6=[datetime.date(2018,12,31)]
lldia7=[datetime.date(2018,3,31),datetime.date(2018,4,1),datetime.date(2018,4,2
),datetime.date(2018,6,2),datetime.date(2018,6,3),datetime.date(2018,6,9),datet
ime.date(2018,6,10),datetime.date(2018,6,16),datetime.date(2018,6,17),datetime.
date(2018,6,23),datetime.date(2018,6,24),datetime.date(2018,6,30),datetime.date
(2018,7,1),]
llpossdia=[lldia0,lldia1,lldia2,lldia3,lldia4,lldia5,lldia6,lldia7]

# Horaris de cada tipus de dia
horari1=[8.5,20.5]
horari2=[8.5,14]
horari3=[9,22]
horari4=[8.5,22]
horari5=[8.5,17]
horari6=[9,14]
horari7=[10.5,20]
llposshorari=[horari1,horari2,horari3,horari4,horari5,horari6,horari7]


def biblioooberta(fecha,horadeldia):
    tipushorari=tipushoraribiblio(fecha)
    res=mirarhorari(horadeldia,tipushorari)
    return [res,tipushorari]

def tipushoraribiblio(data):
    # Per mirar de quina hora a quina hora obra aquell dia la biblio
    # Retorna el tipus de dia que és: A,B,C,D,E,F o G
    tipus=0

    setmanadelany = data.isocalendar()[1]
    if data.year == 2019:
        setmanadelany=setmanadelany+100

    if data.isocalendar()[2] in [6,7]:
        finde=True
    else:
        finde=False

    # Primera seleccio. Mirant si és finde i setmana de l'any
    if not finde:
        n=0
        for llsetmX in llposssetm:
            if setmanadelany in llsetmX:
                tipus=n
                break
            n=n+1

    # Segon filtre mirant si és algun dia concret
    n=0
    for lldiaX in llpossdia:
        if data in lldiaX:
            tipus=n
```

```
            break
        n=n+1
    return tipus

def mirarhorari(horadeldia,tipushorari):
    # Per mirar si a aquella hora està oberta la biblio
    # Retorna un booleà: True si la biblio està oberta

    oberta = False
    if tipushorari!=0:
        [hobra,htanca]=llposshorari[tipushorari-1]
        if horadeldia>hobra and horadeldia<=htanca:
            oberta = True

    return oberta
```

A.3. Funció calendariacademic.py

```
import datetime

def calendariacademic(data,setmanadelany):

    if data.year==2019:
        setmanadelany=setmanadelany+52

    # Periode amb la variable q: 1=preparcial, 2=parcials, 3=postparcials,
4=finals, 5=reavas, 6=nul
    # Variable setmanes fins examens. En cas que estiguem en periode nul, serà
igual a '-'
    setmanesfinsexams='-'
    if data <= datetime.date(2018,6,26):
        q=1000
        parcials = 14
        finals = 23
    elif data <= datetime.date(2018, 7, 6):
        q = 5
        setmanesfinsexams = 0
    elif data <= datetime.date(2018, 9, 11):
        q = 6
    elif data <= datetime.date(2019, 1, 25):
        q = 1000
        parcials = 44
        finals = 54
    elif data <= datetime.date(2019, 2, 10):
        q = 6
    elif data <= datetime.date(2019, 6, 28):
        q = 1000
        parcials = 14+52
        finals = 23+52
    elif data <= datetime.date(2019, 7, 10):
        q = 5
        setmanesfinsexams = 0
```

```
    else:
        q=6

    if q==1000:
        if setmanadelany < parcials:
            q = 1
            setmanesfinsexams=parcials-setmanadelany
        elif setmanadelany == parcials:
            q = 2
            setmanesfinsexams=0
        elif setmanadelany < finals:
            q = 3
            setmanesfinsexams=finals-setmanadelany
        else:
            q = 4
            setmanesfinsexams=0

    return [q,setmanesfinsexams]
```

A.4. Funció eliminarbibliotancada.py

```
finicial='atributs.csv'
ffinal='fitxerfinal.csv'

dfini = open(finicial, 'r')
dffinal = open(ffinal, 'w')

n=0
for e in dfini:
    e = e[:-1]
    if n==0:
        dffinal.write("Data/Hora;Energia activa total;Setmana de l'any;Dia de
l'any;Dia de la setmana;Hora del dia;Periode;Setmanes fins examens;Tipus horari
biblio\n")
    else:
        l = e.split(';')
        if eval(l[6]):
            dffinal.write(l[0] + ';' + l[1] + ';' + l[2] + ';' + l[3] + ';' +
l[4] + ';' + l[5] + ';' + l[7] + ';' + l[8] + ';'+l[9]+'\n')
    n=n+1

dfini.close()
dffinal.close()
```

A.5. Funció taulacomparacio.py

```
import pandas as pd
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.linear_model import ElasticNet
from sklearn.linear_model import Lars
from sklearn.linear_model import LassoLars
```

```
from sklearn.linear_model import OrthogonalMatchingPursuit
from sklearn.linear_model import BayesianRidge
from sklearn.linear_model import ARDRegression
from sklearn.linear_model import SGDRegressor
from sklearn.linear_model import PassiveAggressiveRegressor
from sklearn.linear_model import TheilSenRegressor
from sklearn.linear_model import HuberRegressor
from sklearn.linear_model import RANSACRegressor
from sklearn.kernel_ridge import KernelRidge
from sklearn.svm import SVR
from sklearn.svm import NuSVR
from sklearn.svm import LinearSVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.neighbors import RadiusNeighborsRegressor
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.cross_decomposition import PLSRegression
from sklearn.cross_decomposition import PLSCanonical
from sklearn.cross_decomposition import CCA
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.neural_network import MLPRegressor
from sklearn.preprocessing import Normalizer

# Utilitzarem sempre la mateixa llavor
seed=7

model=
{0:LinearRegression(),1:Ridge(random_state=seed),2:Lasso(random_state=seed),3:E
lasticNet(random_state=seed),4:Lars(),5:LassoLars(),6:OrthogonalMatchingPursuit
(),7:BayesianRidge(),8:ARDRegression(),9:SGDRegressor(random_state=seed),10:Pas
siveAggressiveRegressor(random_state=seed),11:TheilSenRegressor(random_state=se
ed),12:HuberRegressor(),13:RANSACRegressor(random_state=seed),14:KernelRidge(),
15:SVR(),16:NuSVR(),17:LinearSVR(random_state=seed),18:KNeighborsRegressor(),19
:RadiusNeighborsRegressor(),20:GaussianProcessRegressor(),21:PLSRegression(),22
:PLSCanonical(),23:CCA(),24:DecisionTreeRegressor(random_state=seed),25:Bagging
Regressor(random_state=seed),26:RandomForestRegressor(random_state=seed),27:Ext
raTreesRegressor(random_state=seed),28:AdaBoostRegressor(random_state=seed),29:
GradientBoostingRegressor(random_state=seed),30:MLPRegressor(random_state=seed)
}

numerodemodel=0

fitxer='fitxerfinal.csv'
ftaula='comparaciopreprocess.csv'

df = pd.read_csv(fitxer,sep=';')
dftaula = open(ftaula, 'w')

# Eliminem els outlayers
df.drop([7690], inplace=True)

# Variable a predir i atributs
```

```
y=df['Energia activa total']
X=df[["Setmana de l'any","Dia de l'any","Dia de la setmana","Hora del
dia","Periode","Tipus horari biblio"]]

# Train test data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.33,
random_state=seed)
X_train_train, X_train_test, y_train_train, y_train_test =
train_test_split(X_train,y_train, test_size = 0.33, random_state=seed)

transformer = Normalizer()
transformer.fit(X_train_train)
X_train_norm=transformer.transform(X_train_train)
X_test_norm=transformer.transform(X_train_test)

dftaula.write("Numero Model;Model;R2 Sense preproces;R2 Normalitzant;Millor
opcio;Maxim R2\n")
opcions=['Sense preproces','Normalitzacio']

for numerodemodel in range(0,31):
    nommodel=str(model[numerodemodel]).split('(')[0]
    dftaula.write(str(numerodemodel)+';'+nommodel + ';')

    new_mdl = model[numerodemodel]
    new_mdl.fit(X_train_train, y_train_train)
    y_pred = new_mdl.predict(X_train_test)
    r2=r2_score(y_train_test, y_pred)
    dftaula.write(str(round(r2,4))+';')

    new_mdl_norm = model[numerodemodel]
    new_mdl_norm.fit(X_train_norm, y_train_train)
    y_pred_norm = new_mdl_norm.predict(X_test_norm)
    r2_norm= r2_score(y_train_test, y_pred_norm)
    dftaula.write(str(round(r2_norm,4))+';')

    A=[r2,r2_norm]
    millor=opcions[A.index(max(A))]
    dftaula.write(millor+';'+str(round(max(A),4))+'\n')

dftaula.close()

# Canviar punts per comes
dfini = open(ftaula, 'r')
a=dfini.read().replace('.',',')
dfini.close()
dffinal = open(ftaula, 'w')
dffinal.write(a)
dffinal.close()
```

A.6. Funció finalistes.py

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
```

```
from sklearn.linear_model import Lasso
from sklearn.linear_model import ElasticNet
from sklearn.linear_model import Lars
from sklearn.linear_model import LassoLars
from sklearn.linear_model import OrthogonalMatchingPursuit
from sklearn.linear_model import BayesianRidge
from sklearn.linear_model import ARDRegression
from sklearn.linear_model import SGDRegressor
from sklearn.linear_model import PassiveAggressiveRegressor
from sklearn.linear_model import TheilSenRegressor
from sklearn.linear_model import HuberRegressor
from sklearn.linear_model import RANSACRegressor
from sklearn.kernel_ridge import KernelRidge
from sklearn.svm import SVR
from sklearn.svm import NuSVR
from sklearn.svm import LinearSVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.neighbors import RadiusNeighborsRegressor
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.cross_decomposition import PLSRegression
from sklearn.cross_decomposition import PLSCanonical
from sklearn.cross_decomposition import CCA
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.neural_network import MLPRegressor
from sklearn.preprocessing import Normalizer
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_validate
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt

# Utilitzarem sempre la mateixa llavor
seed=7

model=
{0:LinearRegression(),1:Ridge(random_state=seed),2:Lasso(random_state=seed),3:E
lasticNet(random_state=seed),4:Lars(),5:LassoLars(),6:OrthogonalMatchingPursuit
(),7:BayesianRidge(),8:ARDRegression(),9:SGDRegressor(random_state=seed),10:Pas
siveAggressiveRegressor(random_state=seed),11:TheilSenRegressor(random_state=se
ed),12:HuberRegressor(),13:RANSACRegressor(random_state=seed),14:KernelRidge(),
15:SVR(),16:NuSVR(),17:LinearSVR(random_state=seed),18:KNeighborsRegressor(),19
:RadiusNeighborsRegressor(),20:GaussianProcessRegressor(),21:PLSRegression(),22
:PLSCanonical(),23:CCA(),24:DecisionTreeRegressor(random_state=seed),25:Bagging
Regressor(random_state=seed),26:RandomForestRegressor(random_state=seed),27:Ext
raTreesRegressor(random_state=seed),28:AdaBoostRegressor(random_state=seed),29:
GradientBoostingRegressor(random_state=seed),30:MLPRegressor(random_state=seed)
}

numerodemodel=0

fitxer='fitxerfinal.csv'
```

```
df = pd.read_csv(fitxer,sep=';')

# Eliminem els outlayers
df.drop([7690], inplace=True)

# Variable a predir i atributs
y=df['Energia activa total']
X=df[["Setmana de l'any","Dia de l'any","Dia de la setmana","Hora del
dia","Periode","Tipus horari biblio"]]

# Train test data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.33,
random_state=seed)

results=[]
names=[]
msg=[]

for numerodemodel in [16,18,25,26,27]: #finalistes

    nommodel = str(model[numerodemodel]).split('(')[0]

    kfold=KFold(n_splits=5,random_state=seed)
    cv_resultats=cross_val_score(model[numerodemodel], X_train, y_train,
cv=kfold, scoring='r2')
    cv_resultats2 = cross_validate(model[numerodemodel], X_train, y_train,
cv=kfold, scoring='r2')
    results.append(cv_resultats)
    names.append(nommodel)

resume=(nommodel,cv_resultats.mean(),cv_resultats.std(),cv_resultats2['score_ti
me'].mean(),cv_resultats2['score_time'].std())
    msg.append(resume)

print(msg)

fig=plt.figure()
fig.suptitle("Comaparacio d'algoritmes finalistes")
ax=fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(['NuSVR','KNeighbors\nRegressor','Bagging\nRegressor','Rando
mForest\nRegressor','ExtraTrees\nRegressor',])
plt.xlabel('Models')
plt.ylabel('R2')
plt.show()
```

A.7. Funció optimitzacio.py

```
import pandas as pd
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.model_selection import GridSearchCV
```

```
# Utilitzarem sempre la mateixa llavor
seed=7

fitxer='definitiu3.csv'

df = pd.read_csv(fitxer,sep=';')

# Eliminem els outlayers
df.drop([7690], inplace=True)

# Variable a predir i atributs
y=df['Energia activa total']
X=df[["Setmana de l'any","Dia de l'any","Dia de la setmana","Hora del
dia","Periode","Tipus horari biblio"]]

# Train test data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.33,
random_state=seed)

tuned_parameters = {'n_estimators':[5,10,20,30,50],'criterion':['mse','mae']}

clf = GridSearchCV(ExtraTreesRegressor(random_state=seed), tuned_parameters,
cv=5, scoring='r2')
clf.fit(X_train, y_train)

print(clf.best_score_)
print(clf.best_estimator_)

res=clf.cv_results_
taula=pd.DataFrame.from_dict(res,orient='index')
print(taula)

ftaula='optimitzacio.csv'
dftaula = open(ftaula, 'w')

for i in range (len(taula.columns)):
    dftaula.write(str(taula[i])+'\n\n\n')
```

Annex B

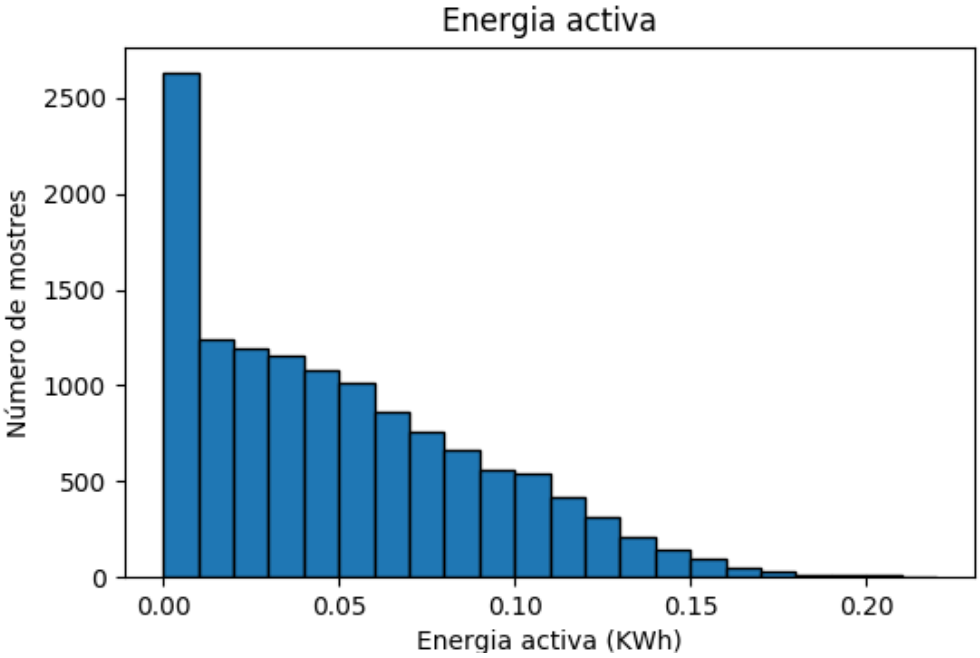

Figura 34. Histograma de les mostres obtingudes d'energia activa

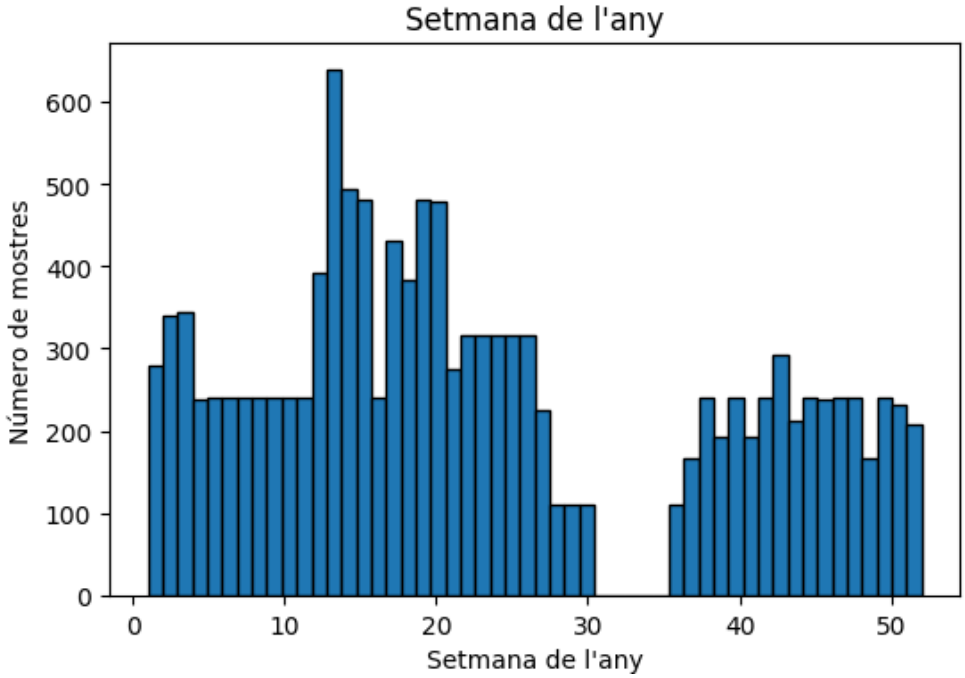

Figura 35. Histograma de les mostres obtingudes depenent de la setmana de l'any

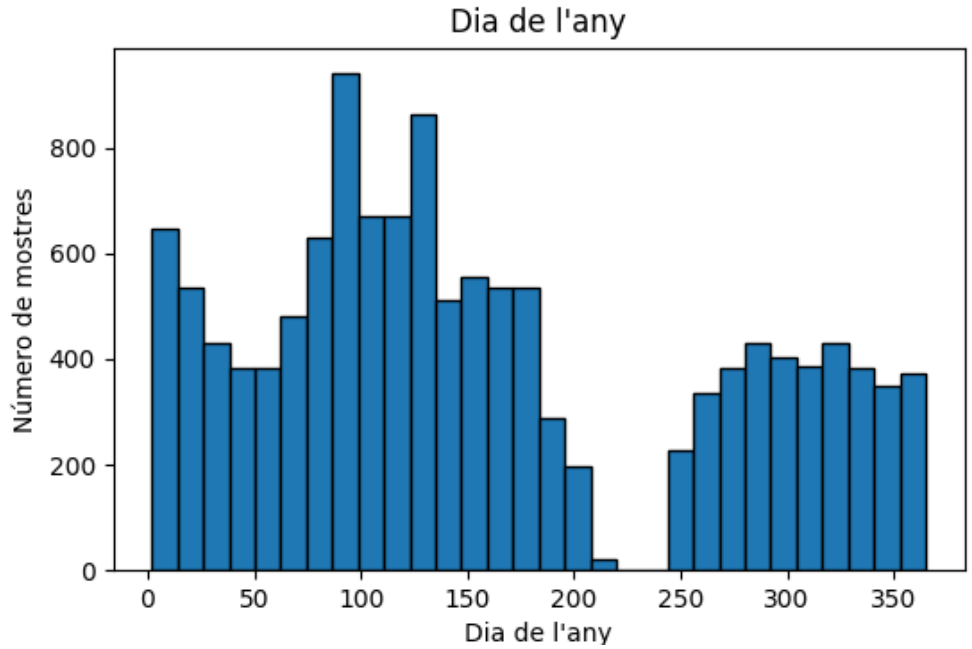

Figura 36. Histograma de les mostres obtingudes depenent del dia de l'any

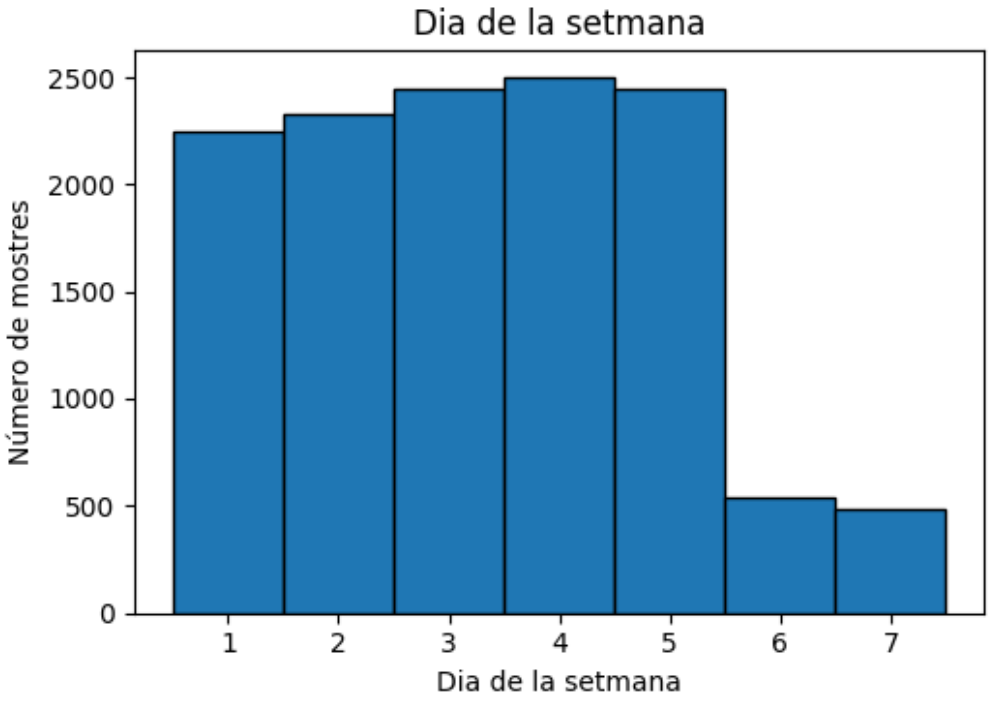

Figura 37. Histograma de les mostres obtingudes depenent del dia de la setmana

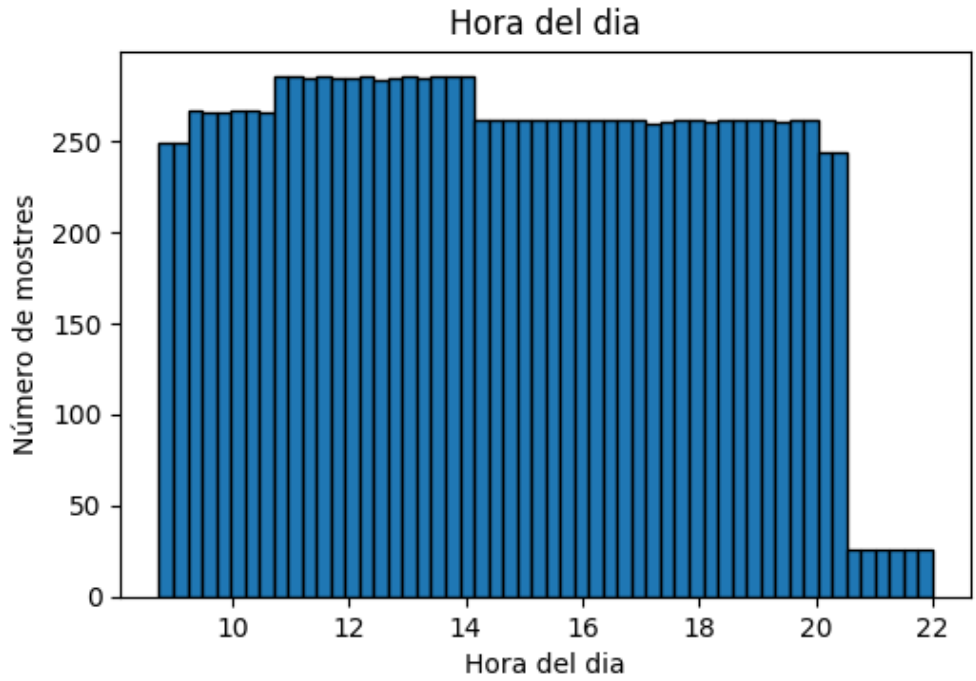

Figura 38. Histograma de les mostres obtingudes depenent de l'hora del dia

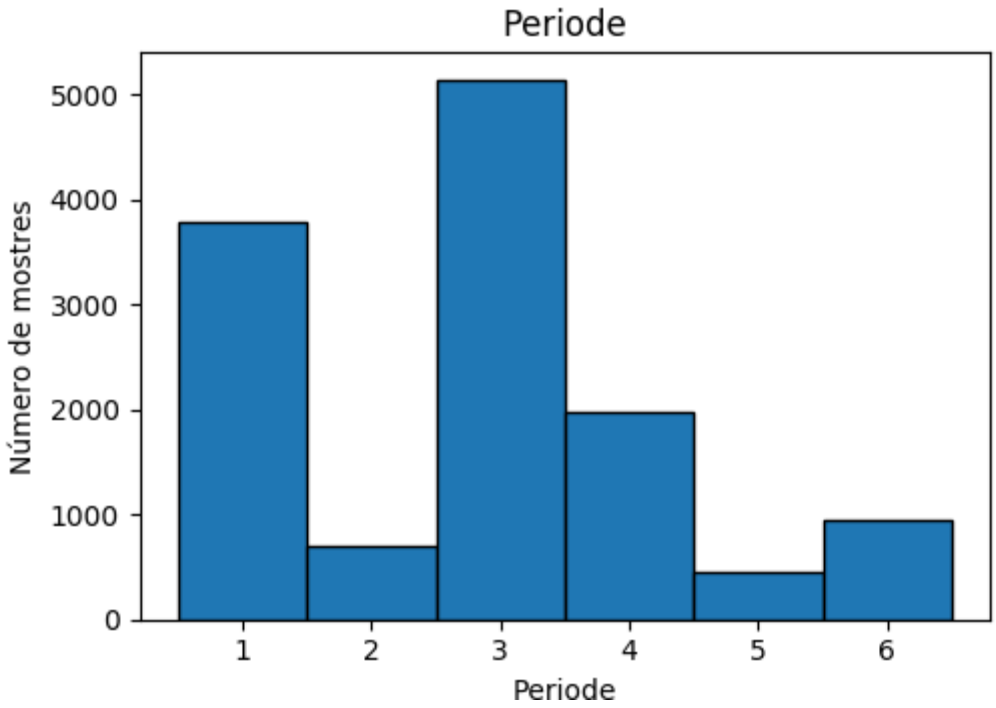

Figura 39. Histograma de les mostres obtingudes per cada període del curs

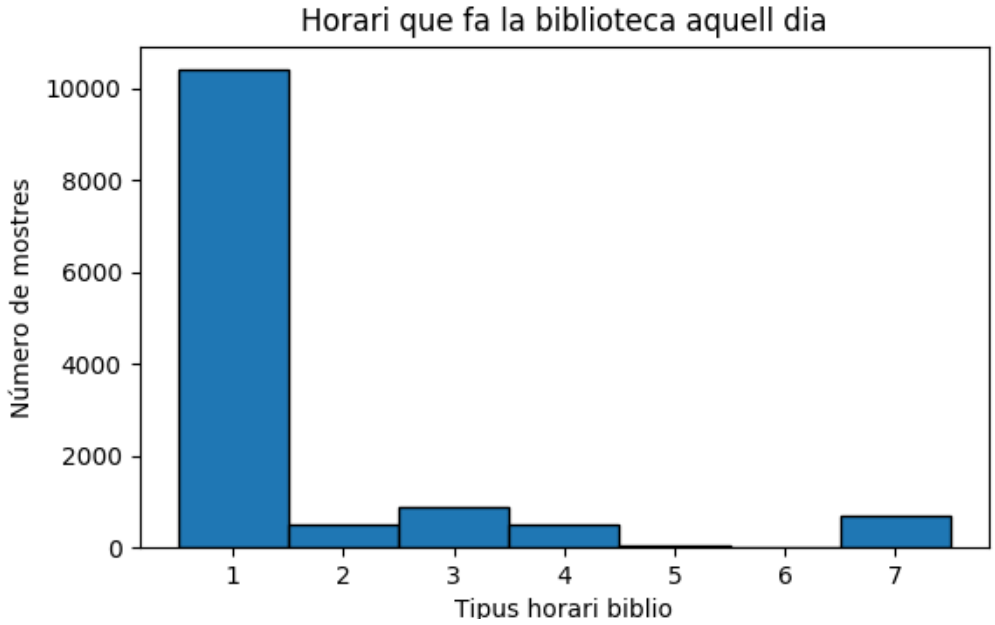

Figura 40. Histograma de les mostres obtingudes per cada horari que fa la biblioteca